Bayesian Models for Multilingual Word Alignment

Robert Östling

Academic dissertation for the Degree of Doctor of Philosophy in Linguistics at Stockholm University to be publicly defended on Friday 22 May 2015 at 13.00 in hörsal 5, hus B, Universitetsvägen 10 B.

Abstract

In this thesis I explore Bayesian models for word alignment, how they can be improved through joint annotation transfer, and how they can be extended to parallel texts in more than two languages. In addition to these general methodological developments, I apply the algorithms to problems from sign language research and linguistic typology.

In the first part of the thesis, I show how Bayesian alignment models estimated with Gibbs sampling are more accurate than previous methods for a range of different languages, particularly for languages with few digital resources available —which is unfortunately the state of the vast majority of languages today. Furthermore, I explore how different variations to the models and learning algorithms affect alignment accuracy.

Then, I show how part-of-speech annotation transfer can be performed jointly with word alignment to improve word alignment accuracy. I apply these models to help annotate the Swedish Sign Language Corpus (SSLC) with part-of-speech tags, and to investigate patterns of polysemy across the languages of the world.

Finally, I present a model for multilingual word alignment which learns an intermediate representation of the text. This model is then used with a massively parallel corpus containing translations of the New Testament, to explore word order features in 1001 languages.

Keywords: *word alignment, parallel text, Bayesian models, MCMC, linguistic typology, sign language, annotation transfer, transfer learning.*

Stockholm 2015
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-115541

ISBN 978-91-7649-151-5

Department of Linguistics

Stockholm University, 106 91 Stockholm

Bayesian Models for Multilingual Word Alignment

Robert Östling

Bayesian Models for Multilingual Word Alignment

Robert Östling

ISBN 978-91-7649-151-5

Printing: Holmbergs, Malmö 2015
Distributor: Publit
Cover: Robert Östling

晓慧，没有你，我该怎么办？
兔娃，你和论文同时孕育，同时诞生，但是论文不如你！

Abstract

In this thesis I explore Bayesian models for word alignment, how they can be improved through joint annotation transfer, and how they can be extended to parallel texts in more than two languages. In addition to these general methodological developments, I apply the algorithms to problems from sign language research and linguistic typology.

In the first part of the thesis, I show how Bayesian alignment models estimated with Gibbs sampling are more accurate than previous methods for a range of different languages, particularly for languages with few digital resources available—which is unfortunately the state of the vast majority of languages today. Furthermore, I explore how different modifications to the models and learning algorithms affect alignment accuracy.

Then, I show how part-of-speech annotation transfer can be performed jointly with word alignment to improve word alignment accuracy. I apply these models to help annotate the Swedish Sign Language Corpus (SSLC) with part-of-speech tags and to investigate patterns of polysemy across the languages of the world.

Finally, I present a model for multilingual word alignment that learns an intermediate representation of the text. This model is then used with a massively parallel corpus containing translations of the New Testament to explore word order features in 1,001 languages.

Contents

Glossary

bitext a parallel text with two languages, see Section 2.1.

colexification the same word is used for expressing different concepts in a language (François 2008), see Section 4.5.

conjugate prior in Bayesian statistics, a distribution H is a conjugate prior for a distribution G if, when applying Bayes' theorem with H as prior and G as likelihood function, the posterior distribution is of the same family as H. See Section 2.4.1.

fertility the number of words in the target language linked to a given word in the source language, see Section 2.3.4.

isolating languages that use only one morpheme per word, as opposed to synthetic languages.

mixing in Markov Chain Monte Carlo (MCMC) models, a measure of how independent each sample is from the preceding samples. Slow mixing could cause a model to be useless in practice, because samples are strongly correlated with the initial value even after a long time.

non-parametric in non-parametric Bayesian modeling, one uses infinite-dimensional distributions where the number of parameters grows with the number of observations, such as the Dirichlet Process (Section 2.4.4) or the Pitman-Yor Process (PYP) (Section 2.4.5).

posterior in Bayesian statistics, a probability distribution representing the belief in a hypothesis after taking some evidence into account, see Section 2.4.1.

precision ratio of identifications that are correct, with respect to some gold standard classification (dependent only on the number of guesses made by the algorithm, not the total number of instances, unlike recall).

prior in Bayesian statistics, a probability distribution representing the belief in a hypothesis before taking some evidence into account, see Section 2.4.1.

recall ratio of correct identifications to the total number of instances, with respect to some gold standard classification (dependent on the total number of instances, unlike precision).

support the support of a distribution is the set of elements with non-zero probability.

synthetic languages that use multiple morphemes per word, as opposed to isolating languages.

token a word token is a particular instance of a word type.

type a word type is an abstract entity that may have a number of instances (tokens).

Acronyms

AER Alignment Error Rate.

CRP Chinese Restaurant Process.

EM Expectation-Maximization.

HMM Hidden Markov Model.

ITG Inverse Transduction Grammar.

MCMC Markov Chain Monte Carlo.

MLE Maximum Likelihood Estimate.

MT Machine Translation.

NLP Natural Language Processing.

NMI Normalized Mutual Information.

NT New Testament.

PoS Part of Speech.

PYCRP Pitman-Yor Chinese Restaurant Process.

PYP Pitman-Yor Process.

SIC Stockholm Internet Corpus.

SMT Statistical Machine Translation.

SSL Swedish Sign Language.

SSLC Swedish Sign Language Corpus.

SUC Stockholm-Umeå Corpus.

WALS World Atlas of Language Structures.

Acknowledgments

This is the easiest part of the thesis to read, but the most difficult to write. Nobody in their right mind is going to be personally offended by, say, a missing word in the evaluation of initialization methods for Bayesian word alignment models (Section 3.4), but publishing an ordered list of everyone you have spent time with during the last five years is a perilous task. For this reason, I will keep the personal matters brief and simply say that I like everyone around me, continuing instead with the people who helped shape this thesis.

My supervisors Mats Wirén, Jörg Tiedemann and Ola Knutsson guided me through five years of uncertainty, bad ideas, distractions, and occasional progress. Towards the end, the thorough review of Joakim Nivre allowed me to smoothen out some of the rough spots content-wise, and Lamont Antieau's proofreading certainly improved the style. Lars Ahrenberg patiently helped me to overcome some confusion on word alignment evaluation methodology. Sharon Goldwater deserves a special mention for providing both the start and end points of this thesis. Östen Dahl and Bernhard Wälchli helped me with many matters related to the New Testament corpus and its application to typology. Further advice on typology was given by Francesca Di Garbo, Calle Börstell and Maria Koptjevskaja Tamm. Francesca Di Garbo also helped with interpreting data on the Romance languages, Yvonne Agbetsoamedo on Sɛlɛɛ, Benjamin Brosig on Mongolian, Calle Börstell, Lars Wallin, Johanna Mesch and Johan Sjons on Swedish Sign Language, Yoko Yamazaki on Japanese, and Britt Hartmann on Swedish.

Tack!

1. Introduction

In this thesis, I approach a number of seemingly very different problems: finding parts of speech in Swedish Sign Language and 1,001 other languages around the world, investigating word order in all those languages, and determining whether or not they make a difference between hands and arms. The common theme that unites these various topics is the method: *word alignment* of parallel texts. This is one of those tasks that seem trivial to the casual observer, and fiendishly difficult to those of us who tried to implement it on a computer. The problem is this: given translated sentences in different languages, mark which words correspond to each other across these languages.

Apart from the applications mentioned above, much of this thesis will be spent developing and exploring the core methods for word alignment, in particular the recent field of word alignment with Bayesian models using MCMC type methods for inference, particularly Gibbs sampling (DeNero et al. 2008; Mermer & Saraçlar 2011; Gal & Blunsom 2013). While previous work has shown MCMC to be an appealing alternative to the Expectation-Maximization (EM) algorithms most commonly used for word alignment, the few existing studies are relatively narrow, and I saw a need for a broader study of Bayesian word alignment models. In a nutshell, Bayesian models make it possible to easily bias the solutions towards what is linguistically plausible, for instance, by discouraging an excessive amount of postulated translation equivalents for each word, or encouraging a realistic frequency distribution of words in the intermediate representation used in the multilingual alignment algorithm of Chapter 5. Given a suitable choice of prior distributions, Gibbs sampling can be easily applied even though the function describing the total probability under the model is very complex.

My main research questions can be expressed in the following way:

1. What are the characteristics of MCMC algorithms for word alignment, how should they be applied in practice, and how do they compare to other methods?

2. How can word alignment be performed in massively parallel corpora comprising hundreds or thousands of different languages?

3. How can word-aligned massively parallel corpora be used to perform investigations in linguistic typology?

4. Can word alignment and annotation transfer be performed jointly in order to improve the accuracy of these tasks?

1.1. How to read this thesis

Very few people will enjoy reading this thesis in its entirety. This is inevitable when several different aspects of linguistics and applied statistics are squeezed into one volume. On a positive note, such diversity means that there is at least *something* here for a wider range of people. What follows is a quick tour of the thesis for a number of different possible readers.

The computational linguist Chapter 2 contains the necessary background for understanding my own contributions, including word alignment models, Bayesian modeling with MCMC inference, and annotation projection. Depending on your specialization, you may want to read selected parts of this chapter. In Chapter 4 I show how word alignment and annotation transfer can be used to benefit each other, and in Chapter 5 I present new models for multilingual word alignment that are scalable to massively parallel texts with hundreds of languages.

The typologist Chapter 4 describes how linguistic annotations can be transferred automatically from one language (such as English) to another (such as the dying Wantoat language of Papua New Guinea, or any of a thousand others) through the use of parallel texts. This kind of transfer, along with the multilingual word alignment methods from Chapter 5, can help giving answers to typological questions. Not perfect answers, necessarily, but answers that are processed more quickly and based on larger samples than could be arrived at without it. I demonstrate this in two case studies, on lexical typology (Section 4.5) and word order typology (Section 5.3).

The sign language researcher Your interest is likely limited to Section 4.3, which describes the use of a transcribed corpus of Swedish Sign Language (SSL) with a translation into written Swedish to transfer part-of-speech tag information from Swedish to SSL, resulting (with manual corrections) in the first sign language corpus automatically annotated with part-of-speech tags.

The computer scientist/statistician If you are not already familiar with Bayesian models, MCMC methods and how they are used in Natural Language Processing (NLP), then Section 2.5.2 on the mathematical background of this study should be of interest, as well as the practical applications of these in Chapter 4 and Chapter 5. This is however not a thesis in mathematics or computer science, so my aim is not to develop new tools or theories in statistics or machine learning.

1.2. Contributions

My contributions in this thesis are of three different types: algorithms, applications and evaluations. These are summarized here, with references to which of the research questions listed above they answer.

1.2.1. Algorithms

The main algorithmic innovations are to be found in Chapter 5, where a method for multilingual word alignment through an interlingua is developed (Question 2), and in Chapter 4, where word alignment and PoS annotation transfer are performed jointly (Question 4).

1.2.2. Applications

Word alignment in itself is not very exciting to most people. With this in mind, I have tried to apply my word alignment algorithms to some problems selected from different areas of linguistics and NLP. Most of these applications would also be possible using methods for word alignment other than the ones I explore, and the purpose is not mainly to test my own algorithms (for that, see below), but to inspire others to use word-aligned parallel texts in their research. One application concerns annotating text (or transcriptions) with part of speech information (Question 4), particularly for languages with few computer-readable resources available. Chapter 4 uses a corpus of 1,142 New Testament translations to explore this, and Section 4.3 contains a more detailed study with Swedish Sign Language. There are also applications from linguistic typology (Question 3), one concerning word order typology (Section 5.3) and another lexical typology (Section 4.5).

1.2.3. Evaluations

Chapters 3 and 4 contain thorough evaluations for a number of different language pairs, and establish the competitiveness of Bayesian word alignment models for a broader set of languages than have been previously explored, as well as providing a solid baseline for the additional experiments performed in these chapters (Question 1). Furthermore, in my experiments I consistently publish the statistics required for computing most of the many word alignment evaluation measures in use. I hope this will create a precedent for others to make results from different studies more easily comparable in the future.

1.3. Relation to other publications

Parts of the work presented in this thesis have been published previously, or are currently in the process of being published. This section gives a list of these publications and their relations to this thesis.

- I use my Stagger PoS tagger throughout Chapter 4 and Chapter 5, since it is the most accurate one available for Swedish and Icelandic. In hindsight, I somewhat regret that I did not use the Icelandic version (developed in cooperation with Hrafn Loftsson) to provide one more source language in the annotation transfer experiments of Chapter 4.
  - Östling, R. (2013). Stagger: An open-source part of speech tagger for Swedish. *North European Journal of Language Technology*, 3, 1–18

  - Loftsson, H. & Östling, R. (2013). Tagging a morphologically complex language using an averaged perceptron tagger: The case of Icelandic. In *Proceedings of the 19th Nordic Conference on Computational Linguistics (NODALIDA 2013)*, NEALT Proceedings Series (pp. 105–119). Oslo, Norway

- Chapter 4 partially overlaps with two articles and one book chapter, the first covering Section 4.3, which has been accepted at the time of writing and should be published by the time this thesis is defended, the next covering Section 4.5, which has been accepted for publication and is undergoing final revisions, and finally an article currently under review that covers Sections 4.1.3 and 4.2:
  - Östling, R., Börstell, C., & Wallin, L. (2015). Enriching the Swedish Sign Language Corpus with part of speech tags using joint Bayesian word alignment and annotation transfer. In *Proceedings of the 20th Nordic Conference on Computational Linguistics (NODALIDA 2015)*. In press
  - Östling, R. (forthcoming). Studying colexification through parallel corpora. In P. Juvonen & M. Koptjevskaja-Tamm (Eds.), *Lexico-Typological Approaches to Semantic Shifts and Motivation Patterns in the Lexicon*. Berlin: De Gruyter Mouton
  - Östling, R. (submitted a). A Bayesian model for joint word alignment and part-of-speech transfer
- The multilingual alignment algorithm from Chapter 5 was briefly summarized in a 2014 article, and the application from Section 5.3 in another article that is presently under review.
  - Östling, R. (2014). Bayesian word alignment for massively parallel texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers* (pp. 123–127). Gothenburg, Sweden: Association for Computational Linguistics
  - Östling, R. (submitted b). Word order typology through multilingual word alignment

2. Background

This chapter is aimed at giving the reader sufficient background knowledge to understand the later chapters of this thesis, where my own results are presented.

Since my work is focused on Bayesian methods for word alignment, this chapter will mainly cover word alignment and relevant models from Bayesian statistics. Chapter 4 uses annotation transfer as part of the word alignment process, so at the end of this chapter there is also an introduction to previous work on annotation projection based on word alignments.

2.1. Parallel texts

A *parallel text* contains translation-equivalent texts in two or more languages. In the most frequently studied case of two languages, this is referred to as a *bitext*—originally a term from translation theory (Harris 1988).

The translation process involved in producing a parallel text can be complicated, and the ideal scenario involving a single source text that has been translated consistently into a number of other languages is often far from reality. De Vries (2007) discusses the long and convoluted translation history of the Bible, which serves as a good introduction to just how complex the creation of a parallel text can be. For reasons of simplicity, most research in NLP takes a more abstract view and does not attempt to model the full complexity of the translation process.

In order to give the reader unfamiliar with the field a sense of the size and characteristics of parallel corpora available to researchers, I will now describe in passing some of the larger parallel corpora commonly used. Figure 2.1 shows the size in both words and number of translations for the following corpora:

Hansards
: The proceedings (*Hansards*) of the Canadian parliament in English and French is an important early parallel corpus, although it contains only two languages.

Europarl
: The Europarl corpus (Koehn 2005) has been used extensively, particularly in the Machine Translation (MT) field. It contains the proceedings of the European Parliament in 21 of the official languages of the European Union.

UN documents
: The United Nations Parallel Text corpus (Franz et al. 2013) contains a large number of United Nations documents in six different languages, totaling roughly 250 million words per language.

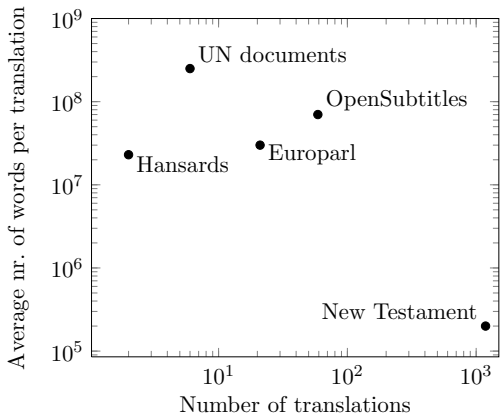


Figure 2.1.: Size in words and number of translations for some important parallel corpora. Note that for the New Testament corpus the number of *translations* is different from the number of *languages*, since some languages have multiple translations. For the other corpora, these figures are equal.

OpenSubtitles
The OpenSubtitles corpus is part of the OPUS corpus collection (Tiedemann 2012) and contains movie subtitles in 59 different languages, translated by volunteers from the OpenSubtitles project.[1]

New Testament
The New Testament of the Christian Bible has been widely translated, and while there is no standard Bible corpus used for NLP research, several authors have used the Bible as a parallel corpus (Cysouw & Wälchli 2007). The corpus used in the present work contains 1,142 translations in 1,001 languages.

The top left corner in Figure 2.1 is occupied by the type of corpora traditionally used in MT and other research, and, as such, contain a small number of languages but a fairly large amount of text for each language. In contrast, the lower right corner contains what Cysouw & Wälchli (2007) term *massively parallel corpora*, here represented by the New Testament. This type of corpus is characterized by a large number of languages, with a relatively small amount of text in each. Massively parallel corpora are less useful for training MT systems, but since they contain a large portion of the roughly 7,000 languages of the world (Lewis et al. 2014) they have been successfully used for studies in linguistic typology (Cysouw & Wälchli 2007).

[1] `http://www.opensubtitles.org/`

2.2. Alignment

In practice, for the type of applications mentioned, one most often needs access to an *aligned* parallel text, where corresponding parts of each translation are matched to each other. This can be done at multiple levels, which are often treated as separate problems and solved using different approaches:

Document linking
: A necessary step before attempting to align a parallel text is to identify which documents in a collection are actually translations of each other. If the required metadata is not available, automatic methods may be applied, although this issue will not be addressed further here.

Sentence alignment
: The first alignment step is typically sentence alignment, where the parallel text is divided into chunks with one or a few sentences per language. Efficient and accurate algorithms for this problem exist, but are beyond the scope of this work. The interested reader is referred to the survey of Tiedemann (2011, chapter 4). In some cases, such as the verses of the New Testament, units other than sentences might be used.

Word alignment
: Given a sentence alignment, the next level of matching is normally done at the word level. Word alignment is a much more challenging problem than sentence alignment and will be discussed further in Section 2.3.

Morpheme alignment
: For synthetic languages, where words may consist of multiple morphemes, word alignment becomes insufficient. Much current research focuses on how to identify and align morphemes rather than whole word forms, which will also be discussed in Section 2.3.

Document linking and sentence alignment are rather well-defined, since (for any reasonable translation) documents and sentences in one language normally have direct correspondences in the other, although sometimes this relationship is not one-to-one. At the word and morpheme levels, on the other hand, the situation is less clear. Several factors speak against even trying to align words or morphemes:

1. words in idiomatic expressions often have no direct counterparts
2. translators can choose different wordings to express roughly the same meaning
3. grammatical morphemes (free or bound) differ widely across languages.

Given this, why even bother trying to align words or morphemes? In spite of all the cases where there is no clear alignment, there are also many examples of the opposite. In most cases, there would be little doubt about which word is used to translate a concrete noun

like *tiger*. Empirically, a large number of applications have demonstrated the usefulness of word (and to some extent morpheme) alignment, but it is important to keep in mind that there are many cases where word alignment is not very well-defined.

Och & Ney (2003) make this uncertainty an integral part of their evaluation metric, the Alignment Error Rate (AER), which assumes that the gold standard annotation data contains three types of relations between source and target language words: *sure link*, *no link* and *possible link*. Although this division and the AER measure in particular have been criticized on the grounds that they correlate poorly with machine translation quality (Fraser & Marcu 2007; Holmqvist & Ahrenberg 2011), this division has been widely used when evaluating word alignment systems. Vilar et al. (2006), on the other hand, argue that AER is an adequate measure of alignment quality as such, but that word alignments are not necessarily a very useful concept for phrase-based machine translation.

It should also be mentioned that given the existence of some very clear alignments, on the one hand, and questionable or even unalignable words, on the other, there is a tradeoff between precision and recall for word alignment. In some applications the recall does not have to be very high, which means that uncertain alignments can be sacrificed in order to improve precision. Liang et al. (2006, figure 2) and Cysouw et al. (2007, figs. 3–5) demonstrate how this tradeoff can look in practice.

The issues surrounding word alignment become even more severe when we move beyond two languages. One way to generalize is by considering each pair of languages, so word w_i^A in language A is aligned to w_j^B in language B, w_k^C in language C, and so on. Beyond the impracticality of quadratic complexity in the number of translations, this also introduces the possibility of inconsistent solutions. For instance, given the A-B and A-C alignments above, what if in the B-C alignment w_j^B was linked to $w_{k'}^C$, for some $k' \neq k$?

An alternative to pairwise alignments is to use a common representation, to which every translation is aligned. I borrow the term *interlingua* from the field of MT to denote this common representation, since the goal is to use a language-independent semantic representation. Because each word is only aligned once there is no risk of inconsistency, and complexity is linear in the number of translations. The question is of course how to find a suitable interlingua, and this problem is dealt with in Chapter 5.

2.3. Word alignment

Most of the research on word alignment focuses on unsupervised algorithms. There are two important reasons for this: lack of annotated data, and the already-acceptable performance of unsupervised algorithms. Given that the world has roughly 7,000 languages, there are around 25 million language *pairs*. Manually word-aligned data that could be used for training supervised word alignment algorithms exists only for an exceedingly small subset of these pairs, whereas massively parallel corpora containing over a thousand languages exist (see Section 2.1) that could readily be used with unsupervised alignment algorithms.

2.3.1. Co-occurrence based models

The word alignment problem is commonly understood as this: for each word in a parallel text, find the word(s)—if any—in the other language(s) that correspond to this word. That is, we want to align words on a token basis.

A somewhat simpler problem is to align word types, so that we can find that e.g. English *dog* tends to correspond more strongly to the German *Hund* 'dog' than to *Katze* 'cat.' This is essentially the problem of lexicon construction, which has received considerable attention on its own (Wu & Xia 1994; Fung & Church 1994).

Given a way of obtaining translational similarity between words types in a parallel text, various heuristics have been tried to transform these into token-level word links (Gale & Church 1991; Melamed 2000). However, in the evaluation of Och & Ney (2003), such methods are much less accurate than the probabilistic IBM models.

Part of this difference can be explained by the fact that co-occurrence based models, unlike the IBM models (except model 1), do not take the structure of sentences into account. Regularities in word order and the fact that words are normally aligned to one or a few other words provide strong constraints on which alignments are likely, so models that ignore this information tend to suffer.

2.3.2. Multilingual word alignment

Most work in word alignment has been carried out on bitexts alignment, where there are exactly two translations. There is however a growing number of parallel corpora with more than two translations, sometimes reaching up to over a thousand translations (see Section 2.1).

Given that bitext alignment is already a quite developed field, perhaps the most natural way of dealing with multilingual parallel texts is to perform pairwise alignment using established methods. Some authors have suggested ways of exploiting additional language for bitext alignment, primarily through various uses of bridge languages. There has also been a separate line of research focusing on aligning word and morpheme *types* in massively parallel texts. We now turn to a discussion of some important examples of approaches to multilingual word alignment.

2.3.2.1. Bridge languages

Several authors have used *bridge languages* (sometimes called *pivot languages*) to improve bitext alignment. The basic idea is that by aligning language A to each of languages X_i ($i = 1, 2, \dots$), and aligning these languages to language B, we obtain information that is useful when aligning A and B.

Borin (2000) uses the union of alignment links discovered through the A–B alignment, on the one hand, and links discovered from the chained A–X_i–B alignments, on the other. He concludes that little precision is lost using this method, while there is a substantial gain in recall.

Kumar et al. (2007) introduce a general probabilistic method for combining word alignments and test it by combining alignments through different bridge languages. Although

AER scores are worse, they show increased Statistical Machine Translation (SMT) performance using their method.[2]

Filali & Bilmes (2005) use a method where the languages of interest, A and B, are in the middle of an alignment chain X–A–B–Y. Given alignments X–A and B–Y, candidate alignments between A and B can be weighted by how reasonable an alignment between X and Y they produce. One way to interpret this is as a kind of inverted bridge language alignment (where the languages of interests *are* the bridge); another is to see the aligned words of X as "tags" to words in A and words in Y as tags to their corresponding words in B. In light of this interpretation, the method is similar to the PoS-enhanced alignment method of Toutanova et al. (2002).

2.3.2.2. Sampling-based alignment

Lardilleux & Lepage (2009) presented a model that learns translation equivalent words and phrases from a multilingual parallel text by sampling small subsets of the text and looking for phrases that occur in the same contexts. Later work (Lardilleux et al. 2011, 2012) improved and extended this model to produce word alignments. Unfortunately neither of these works evaluated word alignment accuracy, but did demonstrate that SMT performance increases if IBM model 4 alignments are combined with their method.

2.3.2.3. Word alignment in massively parallel texts

Massively parallel texts (see Section 2.1) have been used in linguistic typology for automatically comparing some features across a large sample of languages. This specialized type of data (hundreds of languages) and applications (typological investigations) has caused alignment of massively parallel texts to develop mostly independently of SMT-directed word alignment research. While the methods used have so far been rather primitive, the range of applications is impressive and a strong motivation in my own search for improved methods of multilingual word alignment. Examples of this can be found in Cysouw et al. (2007) and Dahl (2007), who use simple co-occurrence statistics for pairwise alignment. The other works in the collection introduced by Cysouw & Wälchli (2007) are also worth consulting for an overview of how massively parallel texts have been applied to crosslinguistic studies.

The only truly multilingual method I am aware of (Mayer & Cysouw 2012) only performs word *type* alignment, carried out by clustering word types from different languages based on co-occurrence statistics. Although both methods and applications are different, this makes it similar to the work of Lardilleux & Lepage (2009).

[2] Kumar et al. (2007) claim in their summary that "Despite its simplicity, the system combination gives improvements in alignment and translation performance," but the improvement in alignment they refer to is relative to the case where there is no bitext between the languages of interest, only indirectly through bridge languages (Shankar Kumar, p.c.).

2.3.3. The IBM models

The IBM models (Brown et al. 1993) have been perhaps the most successful family of bitext word alignment models, and in particular the GIZA++ implementation (Och & Ney 2003) has been frequently used in MT and other applications over the last decade.

Brown et al. (1993) present five models (generally referred to as "model 1" through "model 5"), of which the first two will be discussed here along with the similar Hidden Markov Model (HMM)-based model of Vogel et al. (1996). Models 3, 4 and 5 contain more complex models of sentence structure, but the advantages offered over the HMM model are fairly small, particularly when the latter is extended with a fertility model. Therefore, the higher IBM models will not be described here.

2.3.3.1. Fundamentals

At their core, the IBM models are asymmetric translation models that assume the *target language* (usually denoted $\boldsymbol{f}$, think French or foreign) text is generated by the *source language* (usually denoted $\boldsymbol{e}$, think English) side of a bitext. Target language tokens f_j are assumed to be generated by either one source language token e_i or by the special NULL word, which represents target words that have no correspondence in the source sentence. Alignment variables a_j are used to indicate that f_j was generated by e_{a_j}. This is illustrated in Figure 2.2.

Since the models are used for word alignment rather than actual translation, both $\boldsymbol{e}$ and $\boldsymbol{f}$ are observed quantities; only the alignment $\boldsymbol{a}$ needs to be inferred.

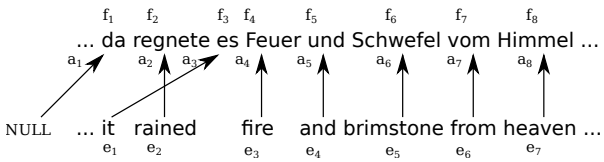


Figure 2.2.: Variables in the IBM alignment models.

All of the IBM models have in common that word correspondences are modeled using categorical distributions conditional on the source token: $p_t(f|e)$. The models differ mainly with respect to how word order is modeled, and whether or not the number of target tokens generated by a single source token is taken into account.

2.3.3.2. Model 1

According to IBM model 1, sentences are generated from the source language sentence $\boldsymbol{e}$ to the target language sentence $\boldsymbol{f}$ in the following way:

1. Choose a target sentence length J from $p_l(J|I)$.
2. For each $j = 1 \ldots J$, choose a_j with probability $p_a(a_j) = 1/I$.

3. For each $j = 1 \dots J$, choose f_j with probability $p_t(f_j|e_{a_j})$.

The probability of a sentence under this model is

$$P(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e}) = p_l(J|I) \prod_{j=1}^{J} p_a(a_j) p_t(f_j|e_{a_j}) \tag{2.1}$$

$$= \frac{p_l(J|I)}{I^J} \prod_{j=1}^{J} p_t(f_j|e_{a_j}) \tag{2.2}$$

which means that all permutations of the target sentence are assumed to be equally likely, and only the translation probabilities matter. If one wants to allow unaligned target words, a NULL token can simply be appended to each source language sentence. In practice, $p_l(J|I)$ is assumed to be constant and does not affect learning alignments.

The assumption that all alignments are equally likely is problematic, since it only holds if word order is arbitrary. In practice, the word order in one language of a bitext tends to be a strong predictor for the word order in the other language. Empirical evaluations have demonstrated that the performance of model 1 is quite poor (Och & Ney 2003).

2.3.3.3. Model 2

Model 2 is identical to model 1, except that it conditions translation alignment links on the position of the target word and the lengths of the source and target sentences, $p_a(i|j, I, J)$. This leads to the following generative story:

1. Choose a target sentence length J from $p_l(J|I)$.
2. For each $j = 1 \dots J$, choose a_j with probability $p_a(a_j|j, I, J)$.
3. For each $j = 1 \dots J$, choose f_j with probability $p_t(f_j|e_{a_j})$.

The total probability of a sentence then becomes

$$P(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e}) = p_l(J|I) \prod_{j=1}^{J} p_a(a_j|j, I, J) p_t(f_j|e_{a_j}) \tag{2.3}$$

While this at least provides a rudimentary model of word order, conditioning on the absolute position and the lengths of both sentences introduces problems due to data sparsity. For this reason, several authors have explored variants of model 2 with word order models that use fewer parameters (Och & Ney 2003; Dyer et al. 2013).

2.3.3.4. HMM model

While not one of the original IBM models, the HMM model of Vogel et al. (1996) is frequently used in combination with them as a natural extension of model 2. Instead of using a distribution of absolute target location, the HMM model uses a distribution

over the relative position with respect to the previous word's alignment, $p_j(m|I)$, where $m = a_j - a_{j-1}$. The probability under this model is

$$P(\boldsymbol{f}, \boldsymbol{a}|\boldsymbol{e}) = p_l(J|I) \prod_{j=1}^{J} p_j(a_j - a_{j-1}|I) p_t(f_j|e_{a_j}) \tag{2.4}$$

This helps to model the fact that sentences do not tend to be reordered at the word level, but rather at e.g. the phrase level. At phrase boundaries, there may be long jumps, but within phrases the jumps tend to be short. Model 2, with its assumption of independence between a_j for different j is unable to capture this. Evaluations have shown the HMM model to be much better than models 1 and 2, somewhat better than model 3, and somewhat worse than models 4 and 5 (Och & Ney 2003; Gal & Blunsom 2013). Extensions to the HMM model have been developed that further improve performance, to the point of rivaling the best of the IBM models (Toutanova et al. 2002).

2.3.4. Fertility

The IBM models and derived models all assume one-to-many alignment links, where a source word can generate zero or more target words. There is a great amount of regularity to be exploited in how many target words a given source word generates (its *fertility*), and different authors have devised a wide variety of methods to do this.

For instance, if the German *Unabhängigkeitserklärung* 'declaration of independence' is aligned to the three English words *declaration of independence* in a particular instance, the fertility of the token *Unabhängigkeitserklärung* is 3. In another instance, it might be aligned to *independence declaration*, and the fertility then is 2. In a good German-English alignment, we can expect *Unabhängigkeitserklärung* to have a fertility of 2 or 3 often, whereas other values are highly unlikely. Most models with fertility tend to include some distribution $p_f(\phi|e)$ conditioned on the source word, so that the fertility of each word can be learned.

Previously, IBM models 3–5 (Brown et al. 1993) used fertility parameters from categorical distributions, Toutanova et al. (2002) extended the HMM model with a fertility-like parameter, and Zhao & Gildea (2010) used a Poisson-distributed fertility parameter. Gal & Blunsom (2013) mostly followed the original IBM models, but used hierarchical Pitman-Yor priors (see Section 2.4.6) for the fertility distributions. In all cases, a sizeable improvement in alignment quality over the corresponding non-fertility model was reported.

2.3.5. Structural constraints

The alignment models described so far only model very general and language-agnostic aspects of language, such as word correspondences and elementary representations of word order. In some cases, further analysis of the text may be available, which could provide valuable information for an alignment model. Some models have been proposed using either PoS tags or syntactic parses to guide word alignment.

2.3.5.1. PoS tags

Toutanova et al. (2002) used PoS-annotated bitexts and simply introduced an additional PoS translation factor $p_p(t_f|t_e)$ conditioning the target tag (t_f) on the source tag (t_e), into the HMM model of Vogel et al. (1996). They obtained improved word alignments, although the gain over the baselines (IBM model 4 and the HMM model) diminished with increasing training data size.

2.3.5.2. ITGs and syntactic parses

Inverse Transduction Grammars (ITGs) (Wu 1997) assume that both versions of a parallel sentence are generated from the same abstract parse tree, where the order of constituents may be reversed in one language in productions.

Yamada & Knight (2001) presented a model which assumes that the target language sentence is generated from the phrase structure parse tree of the source sentence through a series of transformations. They reported better performance than IBM model 5 on a small English-Japanese alignment task.

Cherry & Lin used a dependency parse tree for one of the languages in a bitext to define alignment constraints (Cherry & Lin 2003; Lin & Cherry 2003b) based on *phrasal cohesion* (Fox 2002). This constraint was later used as a feature in a supervised discriminative alignment model (Cherry & Lin 2006b), and by Wang & Zong (2013) who used it in an unsupervised generative model.

Cherry & Lin (2006a) compared the constraints provided by ITGs and phrasal cohesion, finding that ITG constraints were less rarely broken, although when they used the constraints for selecting among alignments in a co-occurrence based model it turned out that the constraint based on phrasal cohesion made up for this by rejecting more incorrect hypotheses, leading to a higher F_1 score. Further gains (though minor) can be obtained by combining the two constraints. This result is somewhat at odds with an earlier study by Zhang & Gildea (2004), who found that ITG-aided alignment was more accurate than the method of Yamada & Knight (2001), which was based on phrase structure grammar. Cherry & Lin (2006a) explained this divergence by the fact that their own method used the same alignment model, just with different constraints, whereas the models compared by Yamada & Knight (2001) differ considerably in other ways.

Lopez & Resnik (2005), on the other hand, evaluated an extension to the HMM model (Vogel et al. 1996) that considers the dependency tree distance between aligned tokens, but this did not lead to any improvement in their evaluation using three different language pairs.

2.3.6. Stemming and lemmatization

Given that data sparsity is a particularly serious problem for word alignment (or indeed any NLP task) with synthetic languages, several researchers have used stemming, lemmatization or similar techniques to normalize the large number of word forms present in such languages. Ideally, a full lemmatization where the canonical citation form of each token is identified would be available, and Bojar & Prokopová (2006) have shown that

the error rate can be cut in half in Czech-English word alignment by using lemmas rather than full word forms.

Accurate lemmatization software is however only available for a small number of languages, and less accurate techniques have also been used. One possibility is to use annotation transfer so that it is sufficient with a lemmatizer or stemmer for one of the bitext languages, using e.g. the method of Yarowsky et al. (2001). However, Fraser & Marcu (2005) found that a very simple method works even better (for some suffixing languages): cutting off all but the first four letters of each word.

2.3.7. EM inference

Brown et al. (1993) and most of their successors used the EM algorithm (Dempster et al. 1977) to learn the parameters of the IBM alignment models. The EM algorithm finds locally optimal parameters of a latent variable model by iterating the following two steps:

- **Expectation**: Compute the expected values of the latent variables given the current parameters and the observed variables.
- **Maximization**: Update the parameters using the maximum-likelihood estimate given the observed variables and the expected values of the latent variables.

For IBM models 1 and 2, as well as the HMM model of Vogel et al. (1996), the expected values of the latent variables (that is, the alignment variables a_j) can be computed efficiently and exactly. Model 1 additionally has the appealing property that its likelihood function is convex (Brown et al. 1993), which means that the EM algorithm will always find a global optimum, although this global optimum is usually not unique and a bad initialization can result in finding a poor solution (Toutanova & Galley 2011).

The more elaborate IBM models 3, 4 and 5 do not have any known efficient method of computing the expectations needed for EM; instead, approximate search algorithms are used that require good initial parameter values in order to converge to a reasonable solution. Typically, the IBM models are pipelined so that parameters learned by a simpler model are used to initialize successively more complex models (Brown et al. 1993; Och & Ney 2003).

In the Bayesian versions of the IBM models (described further in Section 2.4), other methods of inference must be used (Section 2.5). These are mainly variational Bayes (Section 2.5.1) and Gibbs sampling (Section 2.5.3).

2.3.8. Symmetrization

The IBM alignment models and their relatives are asymmetric, and normally find very different alignments depending on which direction they are run. For instance, the links found in a German-English alignment would be expected to differ from the corresponding English-German alignment of the same text. The errors made in the two different directions are independent to some extent, so it is standard to perform a *symmetrization*

procedure to combine the information contained in both directions. Formally, we can describe this as a process of generating a set L of links (i, j), where $(i, j) \in L$ indicates that word e_i in the source language and word f_j in the target language are linked. The input to the symmetrization process are source-to-target alignment variables a_j (where f_j is aligned to e_{a_j}) and the corresponding target-to-source variables b_i (where e_i is linked to f_{b_i}). A number of possible solutions will be summarized below, offering different tradeoffs between precision and recall.

2.3.8.1. Union

Include a link if it exists in at least one direction. This benefits recall, at the expense of precision. Formally:

$$L = \{(i, j) \mid a_j = i \vee b_i = j\}$$

2.3.8.2. Intersection

Include a link if it exists in both directions. This benefits precision, at the expense of recall. Formally:

$$L = \{(i, j) \mid a_j = i \wedge b_i = j\}$$

2.3.8.3. Growing

One family of symmetrization methods starts from the intersection and gradually grows the alignment into adjacent unaligned words (Och & Ney 2003, p. 33). Empirical evaluations of different symmetrization methods have been performed by Ayan & Dorr (2006) and, in more detail,by Wu & Wang (2007). Overall, these methods result in a lower precision than the intersection and lower recall than the union, but typically a better AER and F-score than either, indicating a better balance between recall and precision.

Algorithm 1 shows the *grow-diag-final-and* method, given that the NEIGHBORS function returns all eight coordinates surrounding (i, j)—that is, including the diagonals. Several similar versions have been explored, including *grow-final-and* (which excludes the diagonal neighbors) and *grow-diag* (which omits the invocations of FINAL-AND).

2.3.8.4. Soft symmetrization methods

The methods above all use discrete (or *hard*) alignments, assuming that we have fixed values of a_j and b_i. In many cases, including the standard IBM models, we can obtain marginal distributions for these variables: $p(a_j = i)$ and $p(b_i = j)$. Matusov et al. (2004) and Liang et al. (2006) proposed different methods for exploiting these marginal distributions to obtain improved symmetric word alignments. An important advantage compared to using discrete alignments is that the tradeoff between precision and recall can be adjusted through some parameter, which could make probabilistic methods more generally useful since different applications differ in preferring high precision or high recall. In Section 3.2.3, I describe how soft alignments can be used with growing symmetrization methods.

Algorithm 1 The *grow-diag-final-and* symmetrization heuristic.

function GROW-DIAG-FINAL-AND(A_1, A_2)
  ▷ Start with the intersection of A_1 and A_2.
  $L \leftarrow A_1 \cap A_2$
  ▷ Then proceed to add links from the union, in different steps.
  GROW(A_1, A_2)
  FINAL-AND(L, A_1)
  FINAL-AND(L, A_2)
  return L
end function

▷ Add links to unaligned words neighboring current links.
function GROW(A_1,A_2)
  while L changes between iterations **do**
    ▷ Consider all current alignments.
    for all $(i, j) \in L$ **do**
      ▷ Check all neighboring points (i', j') that are also in the union.
      for $(i', j') \in$ NEIGHBORS$(i, j) \cap (A_1 \cup A_2)$ **do**
        ▷ If either $e_{i'}$ or $f_{j'}$ is unaligned, add (i', j') to L.
        if $(\neg\exists k.(i', k) \in L) \vee (\neg\exists k.(k, j') \in L)$ **then**
          $L \leftarrow L \cup \{(i', j')\}$
        end if
      end for
    end for
  end while
end function

▷ Expand L with previously unlinked words from A
function FINAL-AND(A)
  for all $(i, j) \in A$ **do**
    ▷ If either e_i or f_j is unaligned, add (i, j) to L.
    if $(\neg\exists k.(i, k) \in L) \vee (\neg\exists k.(k, j) \in L)$ **then**
      $L \leftarrow L \cup \{(i, j)\}$
    end if
  end for
end function

2.3.9. Discriminative word alignment models

Instead of applying the various heuristics discussed in the previous section, it is possible to use manually created word alignments to train a discriminative model that predicts alignment links based on e.g. asymmetric alignments from the IBM models, as well as other information. It turns out that only a small amount of annotated data is sufficient for attaining high accuracy in this way, and this has led to several studies on discriminative word alignment (Taskar et al. 2005; Ayan & Dorr 2006; Fraser & Marcu 2006; Moore et al. 2006; Liu et al. 2010).

Although in some situations it is reasonable to manually create the necessary word alignments to train a discriminative model, for instance when creating a machine translation system between two languages where proficient human annotators are available, this is clearly not a realistic requirement for, e.g., the New Testament corpus. For this reason, and because generative alignment models are normally used as an essential subcomponent of discriminative models anyway, the topic of discriminative word alignment models falls outside the scope of this thesis.

2.3.10. Morpheme alignment

For isolating languages, where there is a one-to-one correspondence between words and morphemes, standard algorithms for word alignment also perform morpheme alignment. However, few languages are perfectly isolating, and the more synthetic a language is the further separated the two tasks become.

Figure 2.3 illustrates the difference between word and morpheme alignment, using two languages with widely different levels of synthesis: English and West Greenlandic. While both mappings are equivalent on a word level, Figure 2.3b is clearly the more informative.

(a) Word alignment (b) Morpheme alignment

Figure 2.3.: English-West Greenlandic word vs. morpheme alignment. Example from Fortescue (1984, p. 144).

Eyigöz et al. (2013) used the same HMM-based word order model as Vogel et al. (1996), but used lexical translation probabilities at the morpheme level instead of (or, although this decreases performance for most of their evaluation settings, in combination with) at the word level. In their evaluation on Turkish-English alignment, their method produces better AER scores than IBM model 4 or Vogel et al.'s HMM model run on unsegmented text. Unfortunately they only evaluate word alignment quality, not morpheme alignment. This makes it difficult to compare their results to the most obvious baseline for morpheme alignment: treating morphemes as words and using standard

word alignment algorithms.

The model of Eyigöz et al. (2013) assumed that both sides of the input bitext are morpheme-segmented. They used a supervised tool for their experiments, but in principle unsupervised methods for morphology induction can be done. Unsupervised learning of morphology is a field of active research, which is beyond the scope of this thesis. A good and fairly recent review of the field was written by Hammarström & Borin (2011), to which the interested reader is referred. In a sentence, the problem of unsupervised morphology learning can be summarized as unsolved, but important subproblems can be solved accurately enough to be of practical use.

Intuitively, it seems clear that parallel texts provide useful information for morpheme segmentation, for instance, since a bound morpheme in one language can be a free morpheme in another. There has been some work integrating morpheme alignment and morpheme segmentation, but results are not overwhelmingly positive. Naradowsky & Toutanova (2011), generalizing earlier work by Chung & Gildea (2009), found that their model actually performed better with monolingual than bilingual morpheme segmentation, measured with segmentation F_1.

Snyder & Barzilay (2008), on the other hand, find that jointly learning morpheme segmentation for both languages in a bitext works better than monolingual segmentation (also in terms of segmentation F_1), particularly for related languages. Snyder & Barzilay (2008) used a corpus of short, filtered phrases (also used by Naradowsky & Toutanova (2011) for most of their experiments), which makes it difficult to tell how generally applicable their methods are in practice. Since they simultaneously sampled all possible segmentations and alignments for each word, it is questionable whether this approach scales to whole sentences.

2.3.11. Evaluation of bitext alignment

There are two general approaches to evaluating bitext alignments: intrinsic evaluation methods that try to evaluate the quality of the word alignments as such, and extrinsic evaluation methods that evaluate some other task (typically SMT) that uses the word alignments as input. The main disadvantages of intrinsic evaluations are that we need to define some measure of alignment quality, but these do not necessarily capture the qualities we are interested in. Extrinsic evaluations, on the other hand, are by definition task-specific and might be complex and time-consuming to set up. In the following, the main focus will be on intrinsic evaluations.

Since the pioneering work of Och & Ney (2000, 2003) on the evaluation of bitext alignment methods, most bitext evaluations have assumed a gold standard consisting of two types of alignment links, *sure* links (S) and *probable* links (P), where $S \subseteq P$ and both are defined over pairs (i, j) indicating that word i of the source language is aligned to word j of the target language.

Given an alignment A to be evaluated, Mihalcea & Pedersen (2003) define precision separately for sure and probable alignments:

$$p_T(A, T) = \frac{|A \cap T|}{|A|} \tag{2.5}$$

and similarly for the recall:

$$r_T(A,T) = \frac{|A \cap T|}{|T|} \tag{2.6}$$

where T is either P or S. Och & Ney (2003) and others have used *precision* (without further qualification) to mean p_P and *recall* to mean r_S. AER attempts to combine aspects of both, although it does not directly use either recall or precision in its definition:

$$\text{AES}(A,S,P) = \frac{|P \cap A| + |S \cap A|}{|S| + |P|} \tag{2.7}$$

Finally, a weighted F-score has been advocated by some as an alternative to AER, correlating better with the BLEU score (Papineni et al. 2002) in an SMT setup for some values of α:

$$F_\alpha(A,S,P) = \frac{1}{\frac{\alpha}{p_P} + \frac{1-\alpha}{r_S}} \tag{2.8}$$

The precise value of α that best correlates with BLEU varies between different evaluations, Fraser & Marcu (2007) found the range 0.2–0.4 for a number of language pairs and settings; Holmqvist & Ahrenberg (2011) found 0.1–0.6 for different amounts of data in a Swedish-English evaluation.

Note that in most cases, the F-score used weighs the P-precision and S-recall together, although Mihalcea & Pedersen (2003) also defined the balanced ($\alpha = 0.5$) F-scores separately for both S and P links, whose formulas simplify to:

$$F_S = \frac{2 p_S r_S}{p_S + r_S} \tag{2.9}$$

$$F_P = \frac{2 p_P r_P}{p_P + r_P} \tag{2.10}$$

While the (weighted) F-score provides a balance between precision and recall, the AER does not quite do this. It is even possible to construct an example where alignment A has higher (i.e. better) precision and recall than alignment B, but also higher (i.e. worse) AER:[3]

$$\begin{array}{llll} |S| = 26 & |P| = 64 & |A| = 32 & |B| = 62 \\ |A \cap S| = 20 & |A \cap P| = 32 & |B \cap S| = 19 & |B \cap P| = 60 \end{array}$$

which gives

	A	B
Precision	1.00	0.97
Recall	0.77	0.73
$F_{0.5}$	0.87	0.83
AER	0.103	0.102

[3]Thanks to Lars Ahrenberg for pointing this out.

Finally, it is important to keep in mind that when evaluating against a gold standard alignment, the particular guidelines used for annotation (and of course the level of adherence to these guidelines by annotators) makes a crucial difference. Holmqvist & Ahrenberg (2011) compared a number of such guidelines. Since multiple annotators are typically used, there is also a need to combine the links assigned by each. Here, Och & Ney (2003) used the intersection of sure links and union of probable links from different annotators to form the final sets of sure and probable links, respectively. Mihalcea & Pedersen (2003) on the other hand used a final arbitration phase where a consensus was reached, which led them to label *all* links as sure. Another method is to align the smallest possible phrase pairs and generate sure links from linked single-word phrases and probable alignments from all word pairs in linked multi-word phrases (Martin et al. 2003, 2005).

2.4. Bayesian learning

A thorough review of the vast amount of work done within Bayesian statistics is far beyond the scope of the present work. Beyond basic statistics, which will not be covered here, the background essential to understand the methods introduced includes Dirichlet priors, non-parametric priors (particularly the Pitman-Yor Process), as well as MCMC methods used to perform inference. A brief introduction to these topics will be given next, with references to further literature for the interested reader.

2.4.1. Bayes' theorem

The starting point of Bayesian statistics is Bayes' theorem:

$$\begin{aligned} p(h|d) &= \frac{p(d|h)p(h)}{p(d)} \\ &\propto p(d|h)p(h) \end{aligned} \tag{2.11}$$

The theorem states that given a prior probability $p(h)$ of a hypothesis h, and a probability $p(d|h)$ of observing some data d given h, the posterior probability $p(h|d)$ taking d into account is proportional to the product $p(d|h)p(h)$.

Running example In the following discussion, I will use IBM model 1 as a running example. Initially, simplified subproblems will be considered, gradually building up to the Bayesian version of model 1 (Mermer & Saraçlar 2011; Mermer et al. 2013) and its generalization (Gal & Blunsom 2013).

2.4.1.1. Bernoulli distributions

To illustrate Bayes' theorem with a simple discrete distribution, consider the problem of predicting what the probability is that an English sentence contains the word *buy* given that its German translation contains the word *kaufen* 'buy.'

In order to use Bayes' theorem, the prior distribution $p(h)$ and the likelihood function $p(d|h)$ need to be specified. For the present example, h denotes whether *buy* is present in the English version of the sentence, and d whether *kaufen* is in the German. Since there are only two outcomes (present or not present), these are modeled with *Bernoulli distributions*, which contain a single parameter p, such that the probability of the first outcome (present) is p, and that of the second outcome (not present) is $1 - p$. The parameters could be estimated from a corpus, and Table 2.1 gives the relevant statistics from the New Testament. From this we can compute the Maximum Likelihood Estimates

	kaufen	$\neg$*kaufen*	Σ
buy	8	5	**13**
$\neg$*buy*	2	7,942	**7,944**
Σ	**10**	**7,947**	**7,957**

Table 2.1.: Number of verses containing *buy* or *kaufen* in the English (King James) and German (Luther) versions of the New Testament.

(MLEs) of the prior distribution

$$p(buy) = \frac{13}{7957}$$
$$p(\neg buy) = \frac{7944}{7957} = 1 - p(buy)$$

and for the likelihood we have

$$p(kaufen|buy) = \frac{8}{13}$$
$$p(\neg kaufen|buy) = \frac{5}{13} = 1 - p(kaufen|buy)$$
$$p(kaufen|\neg buy) = \frac{2}{7944}$$
$$p(\neg kaufen|\neg buy) = \frac{7942}{7944} = 1 - p(kaufen|\neg buy)$$

By using Bayes' theorem, it is now possible to compute the posterior probability distribution of the English sentence containing *buy* given that the German translation contains *kaufen*:

$$\begin{aligned} p(buy|kaufen) &\propto p(buy)p(kaufen|buy) \\ &= \frac{13}{7957} \cdot \frac{8}{13} \\ &= \frac{8}{7957} \end{aligned}$$

$$\begin{aligned} p(\neg buy|kaufen) &\propto p(\neg buy)p(kaufen|\neg buy) \\ &= \frac{7944}{7957} \cdot \frac{2}{7944} \\ &= \frac{2}{7957} \end{aligned}$$

which after normalization gives

$$\begin{aligned} p(buy|kaufen) &= 0.8 \\ p(\neg buy|kaufen) &= 0.2 \end{aligned}$$

Since the probabilities used were estimated directly from the example in Table 2.1, we can immediately see that this result makes sense, since 8 of the 10 verses containing *kaufen* also contain *buy*.

2.4.1.2. Binomial and beta distributions

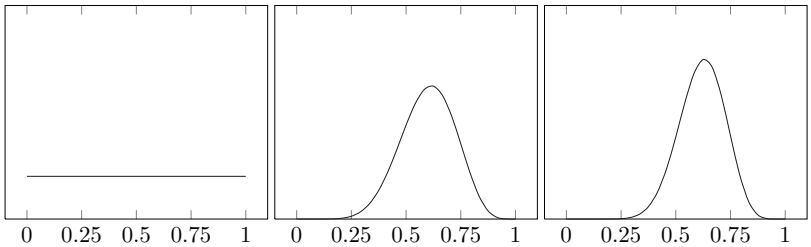
The probabilistic model of translating *buy* looks like this: every time an English sentence with *buy* is translated into German, with probability $p(kaufen|buy)$ add *kaufen* to the translation, otherwise do not.

The value of $p(kaufen|buy)$ is fixed, but unknown. Our only clue is that out of the 13 sentences with *buy* we have seen, 8 contain *kaufen*. In the previous example, we used the following MLE for the data likelihood:

$$\begin{aligned} p(kaufen|buy) &= \frac{8}{13} \\ p(\neg kaufen|buy) &= \frac{5}{13} \end{aligned}$$

Intuitively, it should be clear that the probability is unlikely to be *exactly* 8/13. While the data (8 *kaufen*) is *most* likely given a probability of 8/13, it is almost as likely even if the probability is slightly lower or slightly higher.

We can use Bayes' theorem to model this uncertainty. The hypothesis h in this case is a value $x \in [0, 1]$ representing the underlying probability $p(kaufen|buy)$. The data d contains the information that we have observed $a = 8$ cases of *buy* occurring with *kaufen*, and $b = 5$ cases without *kaufen*.



(a) Prior: Beta$(1, 1)$ (b) 8 of 13: Beta$(9, 6)$ (c) 12 of 19: Beta$(13, 8)$

Figure 2.4.: Prior and posterior distributions $p(h|d)$ of the probability $p(\textit{kaufen}|\textit{buy})$, given a uniform prior and the data that a translations of *buy* contain *kaufen*, while b do not.

The data likelihood is described by a binomial distribution, with success probability x, a successes (translations with *kaufen*) and b failures (translations without *kaufen*):

$$p(d|h) = \binom{a+b}{a} x^a (1-x)^b$$

In order to not favor any particular hypothesis, we use a uniform prior such that $p(h) = 1$ for all $x \in [0, 1]$, otherwise zero.

By Bayes' theorem, we have

$$\begin{aligned} p(h|d) &\propto p(h)p(d|h) \\ &= 1 \times \binom{a+b}{a} x^a (1-x)^b \\ &\propto x^a (1-x)^b \end{aligned}$$

which is proportional to a Beta$(a + 1, b + 1)$ distribution:

$$p(h|d) = \frac{x^a (1-x)^b}{B\big(a+1, b+1\big)}$$

$B\big(x, y\big)$ is the *beta function*

$$B\big(x, y\big) = \frac{\Gamma\big(x\big)\Gamma\big(y\big)}{\Gamma\big(x+y\big)}$$

and the *gamma function* $\Gamma\big(x\big)$ is the generalized factorial function

$$\Gamma\big(x\big) = \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t$$

which as a special case has $\Gamma(n) = (n-1)!$ when $n \in \mathbb{N}$.

The effect on the posterior of adding observations is shown in Figures 2.4b and 2.4c. While some values of $p(\textit{kaufen}|\textit{buy})$ are highly unlikely (below about 0.2, or above about 0.9) the figure shows a rather large amount of uncertainty. Still, compared to the prior (Figure 2.4a) considerable information has been gained by observing that 8 of the 13 sentences with *buy* so far have included *kaufen*.

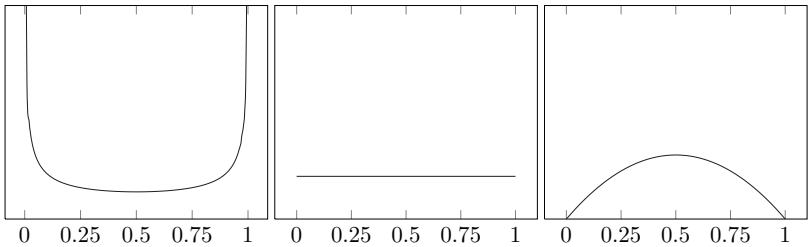
Now, if we want to decrease the uncertainty further we can try to find more parallel sentences with *buy*. Last time we started from a uniform prior, but this time we can do much better by using the posterior distribution from the last step: $\text{Beta}(9, 6)$. Say that we find a new sentences where *buy* is translated into *kaufen*, and b sentences where it is not. Applying Bayes' theorem using this prior, we obtain:

$$\begin{aligned} p(h|d) &\propto p(h)p(d|h) \\ &= \frac{x^8(1-x)^5}{B(9,6)}\binom{a+b}{a}x^a(1-x)^b \\ &\propto x^8(1-x)^5x^a(1-x)^b \\ &= x^{8+a}(1-x)^{5+b} \end{aligned} \tag{2.12}$$

which is a $\text{Beta}(9+a, 6+b)$ distribution. Say that we find six more sentences with *buy*, and four of them turn out to contain *kaufen*. The new posterior then becomes a $\text{Beta}(13, 8)$ distribution, shown in Figure 2.4c. This distribution is more concentrated than the previous posterior (Figure 2.4b), reflecting the decreased uncertainty given the new observations.

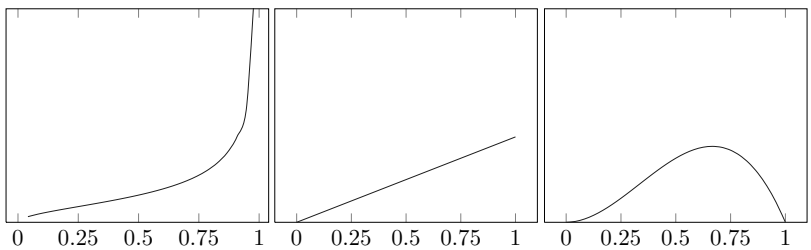
From Equation (2.12), it is easy to see that given a $\text{Beta}(x, y)$ prior on the success probability and an outcome with a successes and b failures from a binomial distribution, the posterior distribution of the success probability is distributed according to $\text{Beta}(x+a, y+b)$. This property makes the beta distribution a conjugate prior of the binomial distribution, a very useful property as we can incrementally add new observations, and the resulting posterior is of the same family of distributions (in this case the beta distribution) as the prior. Noting that the uniform prior we first used is equivalent to a $\text{Beta}(1, 1)$ distribution, Figure 2.4 can be viewed as step-by-step updates of the posterior (always a beta distribution) as more observations are made. Even if we do not have any reason to favor any outcome over the other, the $\text{Beta}(1, 1)$ (uniform) distribution is not the only reasonable option: in general, one can use a $\text{Beta}(\alpha, \alpha)$ distribution for any value $\alpha > 0$. Figure 2.5 shows these distributions for values of α below, equal to, and above 1.

As can be seen in Figure 2.5a, distributions with $\alpha < 1$ assigns higher probability to values near 0 or 1, and less to values around the middle. These distributions generate *sparse* priors and serve to bias the posterior towards distributions where outcomes are either very likely or very unlikely. It would have been reasonable to use a sparse prior instead of a uniform prior in the example above. This would be one way of representing our intuition that either *kaufen* (or any other given word) is the translation of *buy,* or



(a) Beta$(0.5, 0.5)$ (b) Beta$(1, 1)$ (c) Beta$(2, 2)$

Figure 2.5.: Symmetric beta distributions of varying concentration.



(a) Prior: Beta$(0.5, 0.5)$ (b) Prior: Beta$(1, 1)$ (c) Prior: Beta$(2, 2)$

Figure 2.6.: Posterior distributions $p(h|d)$ of the probability $p(\textit{kaufen}|\textit{buy})$, given different priors but the same data: a single occurrence of *buy*, which is translated with *kaufen*.

it is not.[4] The case $\alpha = 1$ is the familiar uniform distribution, and distributions with $\alpha > 1$ are biased *against* extreme points. A prior with $\alpha > 1$ serves to *smoothen* the posterior. As α increases, the posterior becomes more similar to the uniform distribution. Figure 2.6 shows the effect of sparse, uniform and smooth priors when the data consists of a single (positive) example. Since the amount of data is minimal, the prior has a very large effect on the posterior distribution, pulling it towards or away from the MLE ($x = 1$).

2.4.2. The Dirichlet distribution

In the previous section, we considered a simplified translation model where *buy* may or may not be translated into *kaufen* in a given sentence. As described in Section 2.3.3,

[4] Of course, phenomena such as polysemy, synonymy and morphological variants ensure that this assumption does not always hold in reality.

the IBM models assume that the translation of a word e is chosen from a categorical distribution $p(f|e)$.

Fortunately, the methods described in the previous section are straightforward to generalize. The beta distribution (over binomial distributions) is the two-dimensional special case of the Dirichlet distribution, which is a distribution over multinomial distributions.

The parameters $\boldsymbol{\alpha}$ of a d-dimensional Dirichlet distribution are real numbers $\alpha_i > 0$, and the probability density function is given by

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{d} x_i^{\alpha_i - 1}}{B(\boldsymbol{\alpha})}$$

where $B\big(\boldsymbol{\alpha}\big)$ is the multinomial beta function

$$B\big(\boldsymbol{\alpha}\big) = \frac{\prod_{i=1}^{d} \Gamma\big(\alpha_i\big)}{\Gamma\big(\sum_{i=1}^{d} \alpha_i\big)}$$

The Dirichlet distribution is a conjugate prior for the multinomial distribution. Given a prior $h \sim \mathrm{Dir}(\boldsymbol{\alpha})$ and a multinomial observation vector $\boldsymbol{k}$, Bayes' theorem gives

$$\begin{aligned} p(h|\boldsymbol{k}, \boldsymbol{\alpha}) &\propto p(h)p(d|h) \\ &= \frac{\prod_{i=1}^{d} x_i^{\alpha_i - 1}}{B(\boldsymbol{\alpha})} \cdot \frac{\Gamma\big(\sum_i k_i + 1\big)}{\prod_i \Gamma\big(k_i + 1\big)} \prod_{i=1}^{d} x_i^{k_i} \\ &\propto \prod_{i=1}^{d} x_i^{\alpha_i - 1} \prod_{i=1}^{d} x_i^{k_i} \\ &= \prod_{i=1}^{d} x_i^{\alpha_i - 1 + k_i} \\ &\propto \frac{\prod_{i=1}^{d} x_i^{\alpha_i + k_i - 1}}{B\big(\boldsymbol{\alpha} + \boldsymbol{k}\big)} \end{aligned} \tag{2.13}$$

that is, $\boldsymbol{x}|\boldsymbol{k}, \boldsymbol{\alpha} \sim \mathrm{Dir}(\boldsymbol{\alpha} + \boldsymbol{k})$.

Returning to our running example of IBM model 1, we are now ready to model the translation probability distributions using Dirichlet priors. For each source word e, there is a categorical distribution $p(f|e)$ whose parameter is a vector $\boldsymbol{x}$, where x_i is the probability of translating e to f_i.

Observations are vectors $\boldsymbol{k}$, where k_i is the number of times the target word f_i was seen. For instance, *buy* might have been linked eight times with *kaufen*, four times with *kaufe* and once with *anschaffen*. If the German vocabulary looks as follows:

f_1	f_2	f_3	f_4	f_5	...
anrufen	anschaffen	kaufe	kaufen	lachen	...

then this observation would be represented as:

$$\boldsymbol{k} = \big(0 \quad 1 \quad 4 \quad 8 \quad 0 \quad \ldots\big)$$

Before being able to find the posterior distribution of $\boldsymbol{x}$ using Equation (2.13), we also need to decide on a prior. Typically, the priors used are symmetric (because we have no reason to favor particular translations a priori) and sparse (because there is usually only one or a few translations of a given word). A reasonable value for the symmetric prior could be $\alpha = 0.001$, or in other words:

$$\boldsymbol{x}|\alpha \sim \mathrm{Dir}(0.001, 0.001, 0.001, 0.001, 0.001, \dots)$$

Applying Equation (2.13), we obtain the following posterior distribution:

$$\boldsymbol{x}|\boldsymbol{k}, \alpha \sim \mathrm{Dir}(0.001, 1.001, 4.001, 8.001, 0.001, \dots)$$

which means that $\boldsymbol{x}$ is very unlikely to contain probabilities for *anrufen* or *lachen* that are significantly above zero—just what our intuition tells us, given the assumptions that

1. *buy* is likely to be translated into only a few words
2. *buy* has never been observed with *anrufen* or *lachen*.

Note however that, unlike the MLE, this posterior distribution assigns non-zero (although low) probabilities even to distributions that do assert *buy* is likely to be translated into e.g. *anrufen*. Thus if later evidence were to show many instances of *buy* being linked to *anrufen*, the new posterior obtained after taking that information into account would be adjusted to reflect this.

2.4.3. The predictive Dirichlet-categorical distribution

It is often useful to know the probability of one particular outcome k from a categorical distribution, when the parameter vector $\boldsymbol{x}$ of the categorical distribution is itself a random variable distributed according to a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$, and the only known quantities are $\boldsymbol{\alpha}$ and a sequence $\boldsymbol{z}_{1:m}$ of the m first outcomes.

In other words, we want to know the value of $p(z_{m+1} = k|\boldsymbol{\alpha}, \boldsymbol{z}_{1:m})$ given that $\boldsymbol{x} \sim \mathrm{Dir}(\boldsymbol{\alpha})$.

Returning to the example in the previous section, assume that we are interested in modeling how the English *buy* is translated into German. $\boldsymbol{z}_{1:m}$ represents the translations we have seen so far (e.g. $z_1 =$ *kaufen*, $z_2 =$ *kaufe*, $z_3 =$ *kaufen*). From this, and the prior $\boldsymbol{\alpha}$, we would like to know the probability of each word being translated from *buy* next time (that is, the value of z_4).

First, we need the probability of the sequence $\boldsymbol{z}_{1:m}$ given a prior $\boldsymbol{\alpha}$. Let $\boldsymbol{n}$ be a vector such that n_a is the number of times word a occurs in $\boldsymbol{z}_{1:m}$, i.e. $\sum_i \delta_{z_i=a}$. Next, we find $p(\boldsymbol{n}|\boldsymbol{x})$ weighted by $p(\boldsymbol{x}|\boldsymbol{\alpha})$, over *all* different categorical distributions $\boldsymbol{x}$ (where $x_i \geq 0$ and $\sum_i^d x_i = 1$, the d-dimensional probability simplex denoted Δ). This can be obtained in the following way:

$$
\begin{aligned}
p(\boldsymbol{n}|\boldsymbol{\alpha}) &= \int_{\Delta} p(\boldsymbol{n}|\boldsymbol{x}) \cdot p(\boldsymbol{x}|\boldsymbol{\alpha}) \,\mathrm{d}\boldsymbol{x} \\
&= \int_{\Delta} \prod_{i=1}^{d} {x_i}^{n_i} \cdot \frac{1}{B(\boldsymbol{\alpha})} \prod_i {x_i}^{\alpha_i - 1} \,\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{B(\boldsymbol{\alpha})} \int_{\Delta} \prod_{i=1}^{d} {x_i}^{n_i + \alpha_i - 1} \,\mathrm{d}\boldsymbol{x} \\
&= \frac{B(\boldsymbol{n} + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}
\end{aligned} \tag{2.14}
$$

where $B(\cdot)$ is the multinomial beta function (see Section 2.4.2). The last step uses the fact that the product $\prod_{i=1}^{d} {x_i}^{k_i + \alpha_i - 1}$ is simply an unnormalized Dirichlet distribution, so we know that its integral over the probability simplex Δ is equal to the Dirichlet normalization constant $B(\boldsymbol{k} + \boldsymbol{\alpha})$.

Now we can compute the predictive distribution, by using Equation (2.14) in the second step and the fact that $\Gamma(x+1)/\Gamma(x) = x$ in the last step:

$$
\begin{aligned}
p(z_{m+1} = k|\boldsymbol{\alpha}, \boldsymbol{z}_{1:m}) &= \frac{p(\boldsymbol{z}_{1:m}, z_{m+1}|\boldsymbol{\alpha})}{p(\boldsymbol{z}_{1:m}|\boldsymbol{\alpha})} \\
&= \frac{p(\boldsymbol{n} + \hat{\boldsymbol{i}}_k|\boldsymbol{\alpha})}{p(\boldsymbol{n}|\boldsymbol{\alpha})} \\
&= \frac{\Gamma\left(\sum_i (\alpha_i + n_i)\right)}{\Gamma\left(\sum_i (\alpha_i + n_i) + 1\right)} \cdot \prod_i \frac{\Gamma(\alpha_i + n_i + \delta_{ik})}{\Gamma(\alpha_i + n_i)} \\
&= \frac{\alpha_k + n_k}{\sum_i (\alpha_i + n_i)}
\end{aligned} \tag{2.15}
$$

where $\hat{\boldsymbol{i}}_k$ is the kth unit vector (representing a count of 1 for word k, since $z_{m+1} = k$) and δ_{ij} is the Kronecker delta function, defined as

$$
\delta_{ij} = \begin{cases} 0 & \text{when } i \neq j \\ 1 & \text{when } i = j \end{cases}
$$

It is worth reflecting a moment on this simple expression that we had to walk through integrals and long products of gamma functions to arrive at. Equation (2.15) is equivalent to *additive smoothing*, a technique that has long been used to handle unseen events when estimating categorical distributions from observed data.

At a glance, and without being aware of the above, additive smoothing might seem like an ad-hoc method. The intuition behind it is that we can pretend that each k was observed $n_k + a_k$ times (for a smoothing constant a_k) and computing the MLE given these statistics.

On a practical level, the simple form of Equation (2.15) enables very efficient Gibbs sampling in models based on categorical distributions with Dirichlet priors. This is discussed further in Section 2.5.3.

2.4.4. The Dirichlet Process

A Dirichlet distribution is defined over multinomial distributions with a fixed number of d outcomes. In many cases, however, the number of outcomes is not given in advance. This is when non-parametric models are useful, since they can model observations from infinite-dimensional spaces using a finite set of parameters that grows with the number of observations.

The Dirichlet Process, like the Dirichlet distribution, defines a distribution over multinomial distributions. The multinomial distributions drawn from a Dirichlet Process can have countably infinite support, in contrast to the finite-dimensional draws from a Dirichlet distribution. In practice, these multinomial distributions are not sampled directly but are marginalized out so that we can sample multinomial outcomes directly. This requires only a finite number of parameters, which is not given in advance but grows with the number of samples produced.

A Dirichlet Process $\mathrm{DP}(\alpha, G)$ uses a concentration parameter α (analogous to the parameter of a symmetric Dirichlet distribution) and a base distribution G with possibly infinite support. Given previous samples $\boldsymbol{z}_{1:m}$ from a Dirichlet Process, the next value can be sampled using the following distribution:

$$p(z_{m+1} = k | \boldsymbol{z}_{1:m}, \alpha, G) = \frac{n_k + \alpha G(k)}{m + \alpha} \tag{2.16}$$

where n_k is the number of times the value k has been sampled so far, and $\sum_i n_i = m$. Equation (2.16), like the corresponding expression for the Dirichlet distribution in Equation (2.15), is efficient to compute, and this makes the Dirichlet Process and its generalization, the Pitman-Yor Process, popular choices as priors for categorical distributions in Bayesian modeling.

A Dirichlet Process can equivalently be represented as consisting of two parts, a *generator* and an *adaptor* (Goldwater et al. 2006, 2011), providing a more intuitive way of viewing the sampling process. The adaptor in the case of the Dirichlet Process is a Chinese Restaurant Process (CRP), whose name stems from an analogy of a Chinese restaurant with an infinite number of tables. Initially the restaurant has no customers, and each new customer who enters the restaurant decides to sit at an empty table with probability $\alpha/(m + \alpha)$, where m is the number of previous customers, and otherwise chooses another customer at random and sits down next to her. Nobody ever leaves the restaurant. Formally, this amounts to sampling the table assignment z_{m+1} of customer $m + 1$ according to the following distribution:

$$p(z_{m+1} = k | \boldsymbol{z}_{1:m}, \alpha) = \frac{1}{m + \alpha} \cdot \begin{cases} \alpha & \text{if } k = |Z| + 1 \\ n_k & \text{if } 1 \leq k \leq |Z| \\ 0 & \text{otherwise} \end{cases} \tag{2.17}$$

where $|Z|$ is the number of occupied tables (or, equivalently, the number of unique values in $\boldsymbol{z}$), and n_k is the number of customers seated at table k. The likelihood of a seating

arrangement under a CRP is

$$p(\boldsymbol{z}_{1:m}|\alpha) = \frac{\Gamma(1+\alpha)}{\Gamma(m+\alpha)}\alpha^{|Z|-1}\prod_{k=1}^{|Z|}(n_k-1)! \tag{2.18}$$

An important property of the CRP is *exchangeability* between the customers, since the ordering in $\boldsymbol{z}$ does not matter, only the counts n_k of the number of customers at each table. The analogy of the Chinese restaurant can be extended by imagining that whenever a new table is opened, a dish is chosen for that table from a distribution G over dishes. In the terminology of Goldwater et al. (2006), G is the *adaptor*, and they showed that the distribution over the number of customers per dish is equivalent to a Dirichlet Process with concentration parameter α and base distribution G. In NLP, the dishes are typically made to represent word types, while customers represent word tokens.

2.4.5. The Pitman-Yor Process

The CRP can be generalized to the Pitman-Yor Chinese Restaurant Process (PYCRP) by introducing a *discount parameter* d that is subtracted from each table count, which changes the CRP sampling distribution (2.17) to the following:

$$p(z_{m+1}=k|\boldsymbol{z}_{1:m},d,\alpha) = \frac{1}{m+\alpha}\cdot\begin{cases} d\,|Z|+\alpha & \text{if } k=|Z|+1 \\ n_k-d & \text{if } 1\le k\le|Z| \\ 0 & \text{otherwise}\end{cases} \tag{2.19}$$

By replacing the CRP adaptor of the Dirichlet Process with the PYCRP, the Pitman-Yor Process (Pitman & Yor 1997) is obtained. The sampling distribution of $\mathrm{PYP}(d,\alpha,G)$ is

$$p(z_{m+1}=k|\boldsymbol{z}_{1:m},\alpha,G) = \frac{n_k-d\,|Z|+(d\,|Z|+\alpha)\,G(k)}{m+\alpha} \tag{2.20}$$

which as a special case (with $d=0$) gives the corresponding Dirichlet Process sampling distribution (2.16). The counts n_k generated by the CRP and the PYCRP both follow power-law distributions such that $P(n_k=x)\propto x^{-(1+d)}$ (Goldwater et al. 2011). Using $d\approx 0.8$ results in a distribution close to the word frequency distribution in natural languages, commonly known as Zipf's Law. The likelihood of a seating arrangement under a PYCRP is

$$p(\boldsymbol{z}_{1:m}|d,\alpha) = \frac{\Gamma(1+\alpha)}{\Gamma(m+\alpha)}\left(\prod_{k=1}^{|Z|-1}(kd+\alpha)\right)\left(\prod_{k=1}^{|Z|}\frac{\Gamma(n_k-d)}{\Gamma(1-d)}\right) \tag{2.21}$$

To obtain the likelihood of $\boldsymbol{z}$ under a PYP with base distribution G, Equation (2.21) is simply multiplied by the probability $G(l_k)$ of the dish l_k at each table k:

$$p(\boldsymbol{z}_{1:m}|d,\alpha,G) = \frac{\Gamma(1+\alpha)}{\Gamma(m+\alpha)}\left(\prod_{k=1}^{|Z|-1}(kd+\alpha)\right)\left(\prod_{k=1}^{|Z|}\frac{\Gamma(n_k-d)}{\Gamma(1-d)}\right)\left(\prod_{k=1}^{|Z|}G(l_k)\right) \tag{2.22}$$

2.4.6. Hierarchical priors

The Pitman-Yor Process and its special case the Dirichlet Process both use a base distribution G, and nothing prevents this base distribution from being another Pitman-Yor Process. In fact, hierarchical Pitman-Yor priors and their special case, the hierarchical Dirichlet Process (Teh et al. 2006), have been used in areas such as language modeling (Teh 2006), where data is typically sparse but specific on one level (long contexts), and rich but vague on another (short contexts).

A simple example can serve to illustrate this. Imagine that we want to model the distribution of words occurring after the phrase "I saw," denoted $H_{\langle I\ saw\rangle}$. This is a categorical distribution over an infinite number of possible strings, so we can use a Pitman-Yor Process prior on it: $\mathrm{PYP}(d, \alpha, G)$, for discount and concentration parameters d and α (not further considered here), and some base distribution G. A very naive language model could let G be uniform, but this would suffer from data sparsity since there is no way of interpolating with information about shorter contexts.

We can however do much better by letting the base distribution be shared among all the various distributions for strings that end in the same way as "I saw." We call this distribution $H_{\langle I\ saw\rangle}$, and its base distribution $H_{\langle saw\rangle}$ is shared with all bigram distributions ending with "saw." Its base distribution $H_\emptyset$ in turn is shared with all other distributions representing one-word contexts. Finally, the base distribution of $H_\emptyset$ could be e.g. a uniform distribution (stating that all unknown words are equally likely) or some distribution that assigns a probability based on how likely a given letter sequence is to be an English word. The predictive model for the context "I saw" can then be expressed in the following way:

$$
\begin{aligned}
w|\text{ ``}I\ saw\text{''} &\sim H_{\langle I\ saw\rangle} \\
H_{\langle I\ saw\rangle} &\sim \mathrm{PYP}(d_1, \alpha_1, H_{\langle saw\rangle}) \\
H_{\langle saw\rangle} &\sim \mathrm{PYP}(d_2, \alpha_2, H_\emptyset) \\
H_\emptyset &\sim \mathrm{PYP}(d_3, \alpha_3, U)
\end{aligned}
$$

Hierarchical Pitman-Yor Processes have also been applied to word alignment by Gal & Blunsom (2013), who present versions of the IBM models (see Section 2.3.3) in which the various categorical distributions are assumed to have hierarchical Pitman-Yor priors. In the case of IBM model 1, they assume that each translation probability distribution $p_t(\cdot|e)$ has a separate Pitman-Yor Process prior, and that these priors share a common base distribution. The common base distribution thus models the unconditional frequency of each word in the target language, which is interpolated with the per-word translation distributions. Formally:

$$
\begin{aligned}
f|e &\sim G_e \\
G_e &\sim \mathrm{PYP}(d_1, \alpha_1, G_\emptyset) \\
G_\emptyset &\sim \mathrm{PYP}(d_2, \alpha_2, U)
\end{aligned}
$$

In effect, this provides a bias towards linking to common target words. Whether the use of hierarchical priors actually leads to any improvement in word alignment quality is left

unanswered by Gal & Blunsom (2013), but I explore this question further in Section 3.5. Sampling from a hierarchical Pitman-Yor Process is somewhat more complicated than sampling from a non-hierarchical distribution, but efficient algorithms exist (Blunsom et al. 2009).

2.5. Inference in Bayesian models

As was briefly outlined in Section 2.3.7, the IBM alignment models are generally trained using the EM algorithm. Particularly for the simpler models, this can be done exactly and efficiently, but only when simple categorical distributions are used. In the extended Bayesian versions of the IBM models, outlined in Section 2.4.2 and Section 2.4.5, we are forced to look for alternatives to the EM algorithm. A few such methods will be discussed below, with the main focus put on Gibbs sampling, since that is the method I use for the models developed later in this work.

2.5.1. Variational Bayesian inference

Variational Bayesian methods can be used to find approximate solutions in Bayesian models with latent variables (Beal 2003). Riley & Gildea (2012) use this approach to train IBM models 1 and 2 extended with Dirichlet priors on the translation and distortion distributions (see Section 2.4.2), which turns out to require only a minor modification of the standard EM algorithm that does not add to the computational complexity. As expected, they find that using sparse Dirichlet priors improves the accuracy of both word alignment and downstream SMT. Eyigöz et al. (2013) later extend this method for morpheme alignment, but find that results are sometimes worse than the corresponding EM-based algorithm. Unfortunately, for more complex models (such as the ones explored in my own work), deriving the variational Bayes equations can be quite difficult.

2.5.2. Markov Chain Monte Carlo

MCMC methods can be used to obtain unbiased samples from the posterior distribution of a Bayesian model. In general, this is achieved by starting with some assignment of the latent variables, then in each iteration sampling a new assignment from a distribution conditional on the previous assignment. Given that this sampling satisfies certain conditions, the variable assignments sampled during all iterations will approach the true distribution according to the model.

One very important consequence of the dependence on the previous variable assignment is that adjacent samples are correlated. This means that certain parts of the solution space, though probable, may not be visited for a very long time if they are too distant from the starting assignment. In practice, computational resources may only be enough to produce a small number of samples, and if the correlation between adjacent samples is too strong (the *mixing* is slow) this causes a heavy bias towards the starting point. On the other hand, even if the samples represent only a small region of the likely

solution space, this could still be acceptable, just as local maxima are often accepted in deterministic algorithms such as EM.

MCMC methods in general constitute a large research area, and most of it is beyond the scope of this work. Interested readers are instead encouraged to consult the vast literature available. To start with, the novice might benefit from watching the excellent talks of Iain Murray[5] on MCMC methods to get a good overview of the area. More thorough introductions (in writing) exist by e.g. Andrieu et al. (2003), Besag (2004) and MacKay (2003, ch. 29).

The tutorial by Resnik & Hardisty (2010) provides a good introduction from an NLP perspective, with a thorough derivation of a Gibbs sampler for a document topic model. Knight (2009) has written a more informal introduction to Gibbs sampling in NLP.

2.5.3. Gibbs sampling

MCMC methods differ in how the next latent variable assignment is sampled. In Gibbs sampling, this is done by sampling each variable separately (in a deterministic order) conditional on the current value of all other variables. This simple method turns out to be sufficient to guarantee unbiased samples from the model.

Given a suitable choice of distributions, the Gibbs sampling distribution is often easy to derive. For categorical distributions with Dirichlet priors, Equation (2.15) is used, and for Pitman-Yor Process priors Equation (2.20).

We are now able to account for the Gibbs sampling algorithm for a Bayesian IBM model 1 with symmetric Dirichlet priors on the word translation distributions (Mermer & Saraçlar 2011; Mermer et al. 2013). A detailed description is found in Algorithm 2, but for clarity, a high-level summary follows:

1. Randomly initialize all alignment variables.
2. Repeat the following sampling procedure, which at each time t generates a new sample of the alignment variable vector $\boldsymbol{a}^{(t)}$.
   a) Sample each value a_j in turn from $p(i|\boldsymbol{a}_{-j}, \boldsymbol{e}, \boldsymbol{f})$.
   b) Let $\boldsymbol{a}^{(t)}$ be equal to the current value of $\boldsymbol{a}$.

Note that Algorithm 2 describes a *collapsed* Gibbs sampler, where the categorical word translation distributions are never explicitly sampled, but marginalized out as shown in Equation (2.15). It is possible to explicitly represent the conditional distributions as an additional set of variables $\boldsymbol{\theta}$, where $\theta_{f|e}$ is used for $p_t(f|e)$. In each sampling iteration, we would then sample both the categorical distributions $\boldsymbol{\theta}$ and the alignment variables $\boldsymbol{a}$. I follow previous work in using collapsed Gibbs sampling, since $\boldsymbol{\theta}$ contains a very large number of variables and would slow down the sampling considerably, and nobody has demonstrated any benefit in terms of accuracy from using an explicit sampler for this model. Empirical support for this decision is presented in Section 3.3.

[5] `http://videolectures.net/mlss09uk_murray_mcmc/`

Algorithm 2 Collapsed Gibbs sampling for IBM model 1 with Dirichlet priors.

▷ Set alignment counts to zero, they will be initialized below.
$n_{\cdot,\cdot} \leftarrow 0$
▷ Initialize by aligning all target words to a random source word.
for all $\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{f} \in S$ **do**
  for all $j \leftarrow 1 \dots J$ **do**
    ▷ Link to one of the I source words (sampled uniformly).
    $a_j \sim \text{Uniform}(1 \dots I)$
    ▷ Update the counts to reflect this choice.
    $n_{e_{a_j}, f_j} \leftarrow n_{e_{a_j}, f_j} + 1$
  end for
end for
▷ Main part of the algorithm: produce a series of T samples.
for $t = 1 \dots T$ **do**
  for all $\boldsymbol{a}, \boldsymbol{e}, \boldsymbol{f} \in S$ **do**
    for all $j \leftarrow 1 \dots J$ **do**
      ▷ Remove counts that depend on a_j.
      $n_{e_{a_j}, f_j} \leftarrow n_{e_{a_j}, f_j} - 1$
      ▷ Sample one variable (a_j) conditioned on all others ($\boldsymbol{a}_{-j}$).
      $a_j \sim p(i | \boldsymbol{a}_{-j}, \boldsymbol{e}, \boldsymbol{f}) \propto \frac{n_{e_i, f_j} + \alpha}{\sum_k n_{e_i, f_k} + |F|\alpha}$
      ▷ Update the counts with the new value of a_j.
      $n_{e_{a_j}, f_j} \leftarrow n_{e_{a_j}, f_j} + 1$
    end for
  end for
  ▷ All variables of $\boldsymbol{a}$ have now been sampled.
  $\boldsymbol{a}^{(t)} \leftarrow \boldsymbol{a}$
end for

Gao & Johnson (2008) empirically evaluated different methods of estimation for unsupervised HMM learning. Their conclusion was that results depend heavily on the details of the task. More specifically, they found plain EM to be consistently worse than Bayesian methods, and Gibbs sampling generally to provide better results than variational Bayes, although the latter models tend to converge more quickly. Mermer et al. (2013) confirmed that the same relative performance of EM, variational Bayes and Gibbs sampling is valid for the word alignment task. As for collapsed vs. explicit Gibbs samplers, their results varied depending on the size of the data, the number of hidden states, and the evaluation metric used. Although Gao & Johnson (2008) showed that few general conclusions could be drawn without a problem-specific evaluation, their results at least hinted that collapsed Gibbs sampling is a reasonable method compared to the obvious alternatives.

2.5.4. Simulated annealing

Simulated annealing is a stochastic optimization method closely related to MCMC (Kirkpatrick et al. 1983; Richey 2010). The central idea is to use a *temperature* parameter to control the amount of stochastic noise, which is gradually lowered during optimization. This allows for a wide range of solutions to be explored in the beginning, a range that is slowly decreased so that progressively smaller ranges are considered.

It is straightforward to incorporate this idea into a Gibbs sampler. Rather than sampling directly from $p(x_i|\boldsymbol{x}_{-i})$, we sample from

$$\hat{p}(x_i|\boldsymbol{x}_{-i}) \propto p(x_i|\boldsymbol{x}_{-i})^{1/\tau}$$

where τ is the temperature parameter. Some special cases include $\tau = 1$, which is plain Gibbs sampling, $\tau \to 0$, which corresponds to greedy hill-climbing, and $\tau \to \infty$, which samples x_i from a uniform distribution (ignoring $\boldsymbol{x}_{-i}$).

Simulated annealing is particularly useful for problems where the model structure changes in ways that makes it difficult to use the methods described in Section 2.5.5 to estimate marginals from multiple samples.

2.5.5. Estimation of marginals

MCMC algorithms output a series of samples of the latent variables in a model, where each sample is chosen according to the probability assigned to this variable assignment by the model. What exactly should we do with these samples?

For the sake of concreteness, assume that the model is IBM model 1 with Dirichlet priors (see Section 2.4.2). In this case, the samples are vectors $\boldsymbol{a}^{(t)}$ (for each sample $t = 1 \ldots T$) containing alignment variables a_j. Using these samples, there are mainly two things we are interested in approximating: the most probable alignment $\arg\max_{\boldsymbol{a}} p_M(\boldsymbol{a})$ and the marginal distributions $p_{a_j}(i)$ for each alignment variable a_j. We will now go through a few possible ways to use the samples $\boldsymbol{a}^{(t)}$ to approximate these quantities.

2.5.5.1. Using the last sample

The simplest choice is to use the last sample, $\boldsymbol{a}^{(T)}$, to approximate $\arg\max_{\boldsymbol{a}} p_M(\boldsymbol{a})$. Assuming T is high enough, $\boldsymbol{a}^{(T)}$ is an unbiased sample from $p_M(\boldsymbol{a})$. The probability under the model $p_M(\boldsymbol{a}^{(T)})$ is therefore likely to be high, since low-probability alignments are less likely to be chosen. So if the model is good, then $\boldsymbol{a}^{(T)}$ should at least be a reasonable alignment. While using $\boldsymbol{a}^{(T)}$ might sometimes be the best course of action, for many models it is better to instead approximate the marginal values of the variables of interest.

2.5.5.2. Maximum marginal decoding

We can approximate the marginal distributions for each variable using these samples:

$$\begin{aligned} p_{a_j}(i) &= \mathop{\mathbb{E}}_{p_M}[\delta_{a_j=i}] \\ &= \sum_{\boldsymbol{a}} p_M(\boldsymbol{a})\delta_{a_j=i} && (2.23) \\ &\approx \frac{1}{T}\sum_{t=1}^{T} \delta_{a_j^{(t)}=i} && (2.24) \end{aligned}$$

The approximation of each distribution p_{a_j} is useful for a number of purposes beyond estimating $\arg\max_{\boldsymbol{a}} p_M(\boldsymbol{a})$. For instance, a high variance of a_j indicates that the model is unsure about that particular alignment link, and some applications might want to discard it.

If a single alignment is desired, we choose each a_j to be the mode in our approximation of p_{a_j}. In practice, this can be done by choosing the most common value of $a_j^{(t)}$ among the samples $\boldsymbol{a}^{(t)}$ $(t = 1 \dots T)$. Johnson & Goldwater (2009) showed that this improves performance in their word segmentation task, and my own experiments confirm this for various word alignment models.

MCMC samples are correlated, which means that for low t the value of $\boldsymbol{a}^{(t)}$ is biased towards the initial state $\boldsymbol{a}^{(0)}$. Since the initial state typically has a rather low probability, early samples will be biased towards a poor solution, and we might want to discard them so as to not ruin the estimates of p_{a_j}. This is referred to as *burn-in* and leads to improved solutions under some circumstances.

2.5.5.3. Rao-Blackwellization

Although Equation (2.24) approaches the true marginal distribution Equation (2.23) as the number of samples grows, it is possible to obtain faster-converging estimates of the marginal distributions through applying the Rao-Blackwell theorem (Blackwell 1947). This process of Rao-Blackwellization gives the following expression (Gelfand & Smith 1991):

$$\mathop{\mathbb{E}}_{p_M}[\delta_{a_j=i}] \approx \frac{1}{T}\sum_{t=1}^{T} p(a_j^{(t)} = i | \boldsymbol{a}_{-j}^{(t-1)}) \qquad (2.25)$$

Note that $p(a_j^{(t)} = i|\boldsymbol{a}_{-j}^{(t-1)})$ is just the Gibbs sampling distribution for a_j, so these values are computed for each i in any case.

2.5.6. Hyperparameter sampling

So far, we have ignored the values of the parameters to the Dirichlet distributions and Dirichlet and Pitman-Yor Processes used. There are three fundamental ways of dealing with this issue:

1. Take a value from someone else's paper and use that.
2. Investigate a few values on your own problem, then choose the best.
3. Sample the parameters like all the other latent variables.

Johnson & Goldwater (2009) conducted an empirical evaluation of these methods (or at least items 2 and 3) in a word segmentation model and find a considerable gain in accuracy when hyperparameters were sampled. To sample the hyperparameters we need two things: a prior $p(\alpha)$ on a given parameter value α, and a data likelihood function $p(\boldsymbol{x}|\alpha)$. Then, by Bayes' theorem, the posterior from which we sample α is

$$p(\alpha|\boldsymbol{x}) \propto p(\alpha)p(\boldsymbol{x}|\alpha)$$

The likelihood $p(\boldsymbol{x}|\alpha)$ is given by the model, but we are faced with choosing a prior. This made Knight (2009, ch. 27) remark: "Yes, it's turtles all the way down!" Fortunately, the next turtle (the parameters for the hyperparameter prior: hyperhyperparameters) has a smaller effect on model performance, so we can get away with choosing a prior with a high level of uncertainty. Johnson & Goldwater (2009) used uniform priors for the discount parameters of Pitman-Yor Processes, and Gamma$(10, 0.1)$ priors for the concentration parameters. They found no significant differences in accuracy with other values of the gamma distribution parameters. Gal & Blunsom (2013) then repeated that choice of priors for their word alignment models.

In practice, slice sampling (Neal 2003) has proven to be useful for the task of hyperparameter sampling. Slice sampling only requires that we can evaluate a function $\hat{p}(x)$ proportional to the actual probability density function $p(x)$. For instance, if the parameter d of a PYCRP is to be sampled, Equation (2.22) can be used for $\hat{p}(d) = p(d)p(\boldsymbol{z}_{1:m}|d, \alpha)$, where $p(d)$ is some prior and $\boldsymbol{z}$ and α are fixed while sampling d. An accessible introduction to slice sampling can be found in MacKay (2003, pp 374–378), from which Algorithm 3 is adapted. At a higher level, one iteration of the slice sampling algorithm works as follows, starting from point x:

1. Sample u' uniformly between 0 and $\hat{p}(x)$.
2. Find an interval $x_l < x < x_h$ such that $\hat{p}(x_l) < u'$ and $\hat{p}(x_r) < u'$.
3. Sample x' uniformly from the interval, unless $\hat{p}(x') > u'$ the interval is shrunk and x' is resampled from this smaller interval.

Algorithm 3 Slice sampling of a distribution $p(x) \propto \hat{p}(x)$.

function NEXTSAMPLE(x,w,$\hat{p}$)
  ▷ Uniformly sample a point u' below $\hat{p}(x)$
  $u' \sim \text{Uniform}(0, \hat{p}(x))$
  ▷ Find an interval $[x_l, x_r]$ around x such that $\hat{p}(x_l) \leq u'$ and $\hat{p}(x_r) \leq u'$
  $r \sim \text{Uniform}(0, 1)$
  $x_l \leftarrow x - rw$
  $x_r \leftarrow x + (1 - r)w$
  while $\hat{p}(x_l) > u'$ **do**
    $x_l \leftarrow x_l - w$
  end while
  while $\hat{p}(x_r) > u'$ **do**
    $x_r \leftarrow x_r + w$
  end while
  while not returned **do**
    ▷ Draw a candidate sample x' uniformly from $[x_l, x_r]$
    $x' \sim \text{Uniform}(x_l, x_r)$
    if $\hat{p}(x') > u'$ **then**
      ▷ Accept the sample if $\hat{p}(x') > u'$
      return x'
    else
      ▷ Otherwise, make the interval tighter and try again
      if $x' > x$ **then**
        $x_r \leftarrow x'$
      else
        $x_l \leftarrow x'$
      end if
    end if
  end while
end function

2.6. Annotation projection

Given a parallel text, a subset of whose translations are annotated with some linguistic information (e.g. PoS tags), we can use a word alignment to transfer these annotations to other translations in the parallel text.

2.6.1. Direct projection

Given a word alignment where target language tokens f_j are aligned to source language tokens e_{a_j} and some labeling L_e of the source language tokens, we can define a labeling L_f of the target language such that $L_f(f_j) = L_e(e_{a_j})$. Figure 2.7 illustrates the process of direct projection. For instance, in this case *Feuer* 'fire' (f_4) is aligned to *fire* (e_{a_4}). Given that $L_e(e_{a_4}) = \text{NOUN}$, and using direct projection, we then also assume $L_f(f_4) = L_e(e_{a_4}) = \text{NOUN}$. In this well-chosen example from two closely related languages, both the projected PoS and dependency annotations happen to be correct (except the unaligned *da* 'then').

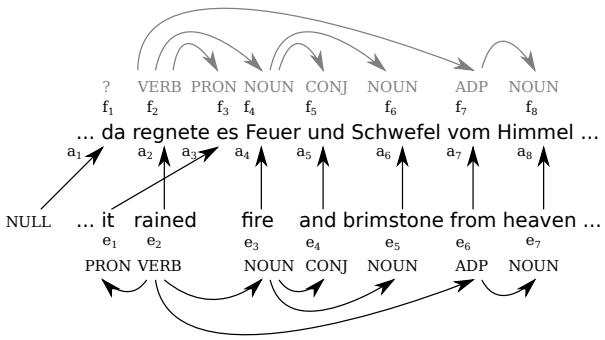


Figure 2.7.: Direct transfer of PoS and dependency annotations from English (below, filled) to German (above, shaded).

Of course, things are not quite that simple in general. We will now go through some of the problems, along with possible remedies. Täckström (2013, section 7.2.1) discusses the problem with direct projection in general, and for more detailed discussions focusing on dependency projections, see Hwa et al. (2002) and Spreyer (2011).

2.6.1.1. Structural differences between languages

There is considerable variation across languages in how words are formed, categorized and combined. Given this variation and the fact that some even argue against the existence of cross-linguistically valid categories (Croft 2001; Haspelmath 2007), it might seem foolish to attempt to transfer, for instance, PoS annotations from one language to another. This is a valid objection in the sense that one should be very careful to draw theoretical conclusions based on automatically projected annotations (and if one is not

convinced by the theoretical arguments, the word alignment error rates are high enough to warrant caution). However, as Figure 2.7 demonstrates, there are clearly cases in which the results of annotation projection are reasonable and useful.

Pragmatically oriented NLP researchers have developed coarse-grained annotation systems for PoS tags (Petrov et al. 2012) and dependency grammar structures (McDonald et al. 2013), designed to be as consistent across languages as possible. While these attempts do not address the underlying theoretical issue of comparability between the structures of different languages, they make great progress towards overcoming a large practical problem, namely that morphosyntactic annotation standards for different languages tend to operate on very different principles, thereby making them more or less incompatible.

2.6.1.2. Errors in word alignments

Automatic word alignment algorithms make errors for a variety of reasons, which were discussed in Section 2.3. Clearly, misaligned words (or unalignable words) can cause additional errors in a direct projection of annotations. Some authors attempt to solve this by filtering out sentences with unreliable alignments, so that the remaining and more accurate parts can be used to estimate parameters for robust target-side models of e.g. PoS tagging (Yarowsky & Ngai 2001; Yarowsky et al. 2001; Hwa et al. 2005; Das & Petrov 2011; Spreyer 2011).

3. Alignment through Gibbs sampling

Although a few works have been published on word alignment through Gibbs sampling, they leave many questions unanswered. The purpose of this chapter is to supplement previous evaluations and to further contribute to the theoretical and practical understanding of Bayesian word alignment models. The bulk of my own innovations will be saved for later chapters, and the present one is restricted to establishing a solid foundation for those methods by filling in some important gaps in previous research.

3.1. Questions

DeNero et al. (2008) presented a non-parametric Bayesian model for phrase alignment intended for phrase-based SMT systems but did not evaluate word-level alignment accuracy. Mermer & Saraçlar (2011) only performed an SMT evaluation with Moses (Koehn et al. 2007), using English-Turkish, English-Czech and English-Arabic corpora. There was no evaluation on manually annotated bitexts, so alignment performance figures are not available. A further limitation of their work is that they only explored IBM Model 1, with Dirichlet priors and fixed hyperparameters. Gal & Blunsom (2013) carried out more extensive experiments on the Chinese-English FBIS corpus[1] evaluating both with an SMT system (also Moses) and directly using AER. As a baseline, they used the GIZA++ implementation (Och & Ney 2003) of the IBM models.

While valuable, these works leave a number of important questions unanswered:

- How do the Bayesian models perform on commonly used word alignment evaluation data sets?
- Is collapsed Gibbs sampling superior to explicit sampling for the word alignment task?
- Do non-parametric hierarchical priors (such as the Pitman-Yor process used by Gal & Blunsom) improve accuracy compared to simple Dirichlet priors (used by Mermer & Saraçlar)?

In this chapter, I attempt to answer these questions in turn.

3.2. Basic evaluations

Since previous studies of word alignment algorithms using Gibbs sampling either have not evaluated alignment performance at all (Mermer & Saraçlar 2011), or have done so

[1] LDC catalog number LDC2003E14, not available for general licensing.

on corpora that are not freely downloadable (Gal & Blunsom 2013), there is a need for evaluations to serve as baselines for future work. In this section, I describe the basic algorithms used and present empirical results for various data sets.

3.2.1. Algorithms

The fundamental building block of this thesis is the collapsed Gibbs sampling algorithm described in Section 2.5.3, where each alignment variable is sampled in turn, conditioned on the current state of all the other alignment variables. There are however many possible variations of and additions to this building block, and I will now go through those that are relevant to the experiments in this chapter.

3.2.1.1. Modeled variables

The technique pioneered by Brown et al. (1993), where simple alignment models are used in a "pipeline" to initialize increasingly complex models, has proved essential in avoiding bad local maxima in complex, non-convex models. While a particular pipeline has been established as standard for the IBM models (models 1 through 5, with model 2 replaced by the HMM model), the variables below could be added in an almost arbitrary order, and multiple variables could be added in a single step. In Section 3.4, I will empirically investigate some of these combinations.

- **Lexical (1)**: In all models, target words f depend on source words e according to the distribution $p_t(f|e)$. Using *only* this dependency results in IBM model 1 (Section 2.3.3.2).
- **Word order (H)**: The HMM alignment model of Vogel et al. (1996), described further in Section 2.3.3.4, adds a distribution $p_j(a_j - a_{j-1}|I)$ describing the length of the "jump" in the source language when moving from token $j-1$ to token j in the target sentence.
- **Fertility (F)**: Next, the number of word aligned to a certain word (its fertility, see Section 2.3.4) is modeled using a distribution $p_f(\phi|e)$.
- **Tags (P)**: When PoS tags are available, these can be modeled in a way analogous to lexical items above, through a distribution $p_p(t_f|t_e)$ representing the dependence of target tags t_f on source tags t_e (Section 2.3.5.1). These tags can be given or transferred between the languages in a bitext, as described in Chapter 4.

I will use the letters in brackets to succinctly describe different alignment models, so that e.g. 1+H+F represents a model with lexical, word order and fertility parameters. In Section 4.1, the letter T will also be used to indicate the PoS transfer algorithm described there.

By assuming conditional independence between these distributions, each distribution simply becomes another factor in the Gibbs sampling distribution for alignment variable a_j.

To make things more concrete, recall that Algorithm 2 shows the sampling process for step 1 above, with only lexical dependencies. When moving to the HMM-based word order model, the only change needed is to the sampling iteration, which for symmetric Dirichlet priors yields:[2]

$$p(i|\boldsymbol{a}_{-j}, \boldsymbol{e}, \boldsymbol{f}) \propto \frac{n_{e_i,f_j} + \alpha_t}{\sum_k n_{e_i,f_k} + |F|\,\alpha_t} \cdot \frac{n_{i-a_{j-1}} + \alpha_j}{\sum_k n_k + N_j\alpha_j} \cdot \frac{n_{a_{j+1}-i} + \alpha_j}{\sum_k n_k + N_j\alpha_j} \tag{3.1}$$

where α_t and α_j are the Dirichlet parameters for the lexical and word order distributions, respectively, and N_j is the number of different jump lengths allowed.

Fertility and tag translation distributions can be added analogously. Note that the computational complexity of a sampling iteration remains quadratic in the length of the sentence, in stark contrast to the EM algorithm, which has cubic complexity for the HMM model and is intractable for the fertility model, forcing implementations to use approximations. This allows the use of longer alignment units than individual sentences, such as Bible verses, or in non-literal translations where it can be difficult to find a good one-to-one sentence alignment.

While Gibbs sampling typically requires a greater number of iterations for acceptable results than EM, the relatively low amount of computation per iteration means that speed is comparable between the approaches.

3.2.1.2. Sampling

Each experiment consists of eight independently initialized models, each producing a long enough series of samples to ensure a very slow change of alignment accuracy. From this, the alignment marginal probabilities are computed in one of two ways:

1. using Rao-Blackwellization (Section 2.5.5.3) from the samples from one of the eight independent models

2. as above, but averaging over all of the independent models.

The purpose of the second method is to cancel out the bias from the random initialization. To illustrate the kind of obstacles a single sampler might face, consider the following example: In the New Testament the words *camel* and *needle* occur mostly together, in the same verses. If one particular random initialization leads to *camel* being aligned multiple times to a word corresponding to *needle* in another language, and vice versa, sampling the opposite (correct) alignment links in a particular sentence will be very unlikely. An even more extreme version of this problem can be found in the explicit sampler explored in Section 3.3. With multiple independent samplers, some are likely to sample mostly in the neighborhood of the correct solution, while others keep near the

[2] Technically, this is incorrect because it assumes the two concurrent draws from the jump length distribution p_j are drawn independently from the same distribution. In practice, the error does not seem large enough to affect the accuracy of the algorithms negatively. For Dirichlet priors an exact solution could be obtained, but to my knowledge this is not possible for a hierarchical PYP prior.

incorrect one. When averaged, the result better reflects the fact that the model assigns relatively high probability to *both* solutions.

Instead of using multiple samplers to circumvent the problems with slow mixing in some regions of the parameter space, one could also try to design a sampler with better mixing properties. In other applications, this has been approached by e.g. type-level sampling (Liang et al. 2010), and it is possible that similar methods could be used to improve Gibbs samplers for word alignment as well. For the HMM model it would also be possible to use sentence-wise blocked sampling, which has turned out to be superior in other HMM-based models (Gao & Johnson 2008).

3.2.2. Measures

Given the problems with existing measures of word alignment quality discussed in Section 2.3.11, as well as the large number of different such measures, I have decided to include the following raw data:

$\|S\|$	number of *sure* alignments in the gold standard
$\|P\|$	number of *probable* alignments in the gold standard
$\|A\|$	number of alignments returned by algorithm
$\|A \cap S\|$	number of algorithm's alignments in the *sure* set
$\|A \cap P\|$	number of algorithm's alignments in the *probable* set

From these figures all of the standard evaluation measures can be derived, including precision, recall, balanced F-scores (combined and for each alignment type) and AER. I will also give some of the standard measures for convenience, despite their being technically redundant.

3.2.3. Symmetrization

The experiments in this section all use the following symmetrization method:

$$A = \{(i, j) \mid p(a_i = j)p(b_j = i) \geq r_I\} \tag{3.2}$$

where r_I is a threshold value, which in this section is fixed at 0.25. This corresponds to a "soft" version of the intersection heuristic described in Section 2.3.8, similar to what was used by Liang et al. (2006). Equivalently, we could express this condition to say that a link (i, j) is added if the geometric mean of $p(a_i = j)$ and $p(b_j = i)$ is above one half. It is of course possible to use other thresholds for different tradeoffs between precision and recall, but the value 0.25 is both theoretically plausible (given the geometric mean interpretation) and empirically strong in my own experience, as well as in the evaluation of Liang et al. (2006, Figure 2). The reason for choosing the soft intersection method is that the algorithms produce rather few NULL alignments and relies on the symmetrization step to exclude alignments that are inconsistent between the two alignment directions.

The growing heuristics described in Section 2.3.8, which start from the intersection and selectively add links from either of the two asymmetric alignments, can be extended

Table 3.1.: Total corpus sizes (in sentences) and number of (S)ure and (P)robable alignment links in their respective evaluation sets.

Corpus	**Sentences**	$\|S\|$	$\|P\|$
WPT-03 English-French	1,130,588	4,038	17,438
WPT-05 Romanian-English	48,641	5,034	5,034
WPT-05 English-Inuktitut	333,185	293	1,972
WPT-05 English-Hindi	3,556	1,409	1,409
Europarl English-Swedish	692,662	3,340	4,577

to the soft case in different ways. The easiest is to first discretize the alignments and then use the standard algorithms. A more flexible method, however, is to compute the intersection using Equation (3.2) and discretize the asymmetric alignments in the following way:

$$\begin{aligned} A_1 &= \{(i,j) \mid p(a_i = j)p(b_j = i) \geq r_S \wedge p(a_i) \geq r_A\} \\ A_2 &= \{(i,j) \mid p(a_i = j)p(b_j = i) \geq r_S \wedge p(b_j) \geq r_A\} \end{aligned} \tag{3.3}$$

This method has three parameters: the thresholds r_I, r_S and r_A. In general, r_I should be fairly high to ensure a high-precision alignment to start with, r_S should be less than r_I so that links do not have to be probable in both directions to be included, and r_A should be high so that links must be probable in at least *one* direction.

This symmetrization method is used in Section 3.5, where the parameters are optimized on annotated development data. Typical values from these experiments are $r_I \approx 0.25$, $r_S \ll 0.1$ and $r_A \approx 0.75$, but this can vary considerably depending on which languages are aligned, and which balance between precision and recall is desired.

3.2.4. Data

I use two different types of corpora in my evaluations: bitext corpora (summarized in Table 3.1) and the New Testament corpus. This section provides brief summaries of the various corpora and references that the interested reader is encouraged to consult. When possible, public data sets that are available free of charge have been chosen.[3]

The shared task of the Workshop on Building and Using Parallel Texts (WPT) (Mihalcea & Pedersen 2003) and its successor in 2005 (Martin et al. 2005) provided several

[3] Rada Mihalcea provides copies of the WPT corpora on her website: `http://web.eecs.umich.edu/~mihalcea/downloads.html` and the Europarl corpus can be found at `http://www.statmt.org/europarl/`. The subset used in my experiments consists of the first 700,000 lines of the sentence-aligned Swedish-English corpus from version 7 of the corpus. Since a few sentences are unaligned, the actual number of parallel sentences used for English-Swedish experiments is slightly lower (see Table 3.1). Finally, the English-Swedish gold standard annotations can be found at `http://www.ida.liu.se/labs/nlplab/ges/`.

hand-aligned data sets for word alignment evaluation, plus unannotated training data. In my evaluations, I used the Romanian-English,[4] English-Inuktitut, English-Hindi and English-French data. Holmqvist & Ahrenberg (2011) created a hand-aligned English-Swedish evaluation set from the Europarl corpus (Koehn 2005). I used this with the first 700,000 sentences of the English-Swedish bitext from Europarl version 7 as training data, to be comparable to the "large" training set of Holmqvist & Ahrenberg.

For some experiments, I also use a corpus of 1,142 New Testament translations into 1,001 languages, as compiled from various Internet sources. A complete list of the languages included can be found in appendix A. Although many of the texts are not covered by copyright, the copyright status of other texts is unclear, so I am unable to make the corpus available through a public website. Since the New Testament corpus contains no explicit alignment links, the method described in Section 5.1.1.4 was used to create gold standard links of the same format as the corpora above.

All of the corpora that were used are sentence-aligned, except for the New Testament corpus, which is aligned at the verse level. Verse alignment is in fact an advantage in a multilingual setting, since it minimizes the impact of translation-specific sentence boundaries.

3.2.5. Baselines

The Bayesian alignment models under consideration require no data except a sentence-aligned bitext, and so comparison should be focused on alignment models operating under similar constraints. However, with further resources it is in some cases possible to improve the accuracy of word alignment algorithms.

3.2.5.1. Manually created word alignments

Although fully supervised word alignment is rare due to the high cost of human annotation of large amounts of data, one line of research uses discriminative models to learn from a small set of manual word alignments plus a larger amount of unaligned bitext (see Section 2.3.9). The obvious disadvantage of models using any kind of supervision is the need for annotated data, which is available for only a handful of language pairs—an exceedingly small fraction of the millions of possible language pairs in the world. This unfortunately prevents their application to e.g. the typological investigations in Section 4.5 and Section 5.3.

3.2.5.2. Other resources

There is a variety of other resources that have been used for word alignment: dictionaries (Och & Ney 2003, p. 32), PoS tags (see Section 2.3.5.1, further explored in Chapter 4), syntactic parses (see Section 2.3.5.2), etc. As was the case with manually created word

[4]Note that both the WPT-03 and WPT-05 shared tasks included Romanian-English, using the same training data but different test sets. The WPT-05 test set appears to be somewhat more difficult, both in my own experiments and in those by Liu et al. (2010, p. 329).

alignments, these resources are only easily available for a small subset of the world's languages.

The WPT shared tasks contained two separate tracks: one for limited resources (unannotated sentence-aligned bitexts, along with small trial sets of manually word-aligned sentence pairs), and one for unlimited external resources (Mihalcea & Pedersen 2003; Martin et al. 2005). Although the trial data sets are small, some researchers have used them for semi-supervised discriminative training and have obtained good results in doing so (Liu et al. 2010, p. 329).

3.2.5.3. Baseline systems

Table 3.2 summarizes the systems used or referred to in this thesis, including my own. The resources required by each system are identified by two independent features.

- **Supervision**: *yes* (requires manually created word alignments) or *no*.
- **Resources**: *both* (requires other resources that are either bilingual or that they must be present for both languages), *one* (requires other resources that must only be present for one language), or *no*.[5]

All of the baseline experiments, except those using GIZA++, have been performed by other authors. While this means that there is considerable diversity in the methods used, the most important point is to find strong baselines to compare against.

Note that several authors have published results using discriminative alignment algorithms on the same datasets I use, but they have generally used part of the test set for discriminative training, which means that the reported figures use different test sets than the original task. Therefore, and because my focus is on unsupervised algorithms, I excluded these from the comparison. One exception is the Vigne system, where an evaluation on the standard test sets of the WPT shared tasks is available (Liu et al. 2010).

As for the GIZA++ experiments, my aim was to simulate the conditions of my own experiments as closely as possible. This meant using no resources except the given bitext, and only default parameters, and the only preprocessing consisted of lowercasing the texts and (for the English-Hindi and English-Romanian experiments only) removing all but the first four characters of each word as a simple kind of stemming (see Section 2.3.6). With GIZA++, the symmetrization method that gives the best result on the test set for each experiment is chosen, in order to provide a baseline that is as strong as possible given the conditions above. Due to limitations in the GIZA++ software, all sentences longer than 100 tokens are discarded. This only affects a very small part of the training set and is not expected to significantly affect the final results. The pipeline used in the

[5] It should be noted that this classification is not always easy. For instance, Fraser & Marcu (2005) optimized one parameter using annotated data for the Romanian-English task they evaluated on and optimize other parameters using annotated data for a closely related language (French). They note "a substantial increase in AER" (p. 92) without this procedure. In questionable cases like this, I give the other author(s) the benefit of the doubt, even though it makes my own results compete against a more unfavorable baseline (see e.g. Table 3.4).

Table 3.2.: Systems compared in this thesis. For the definitions of Sup(ervision) and Res(ources), see the main text.

System	Reference	Sup.	Res.
1+H+F	This work (Section 3.2.1)	no	no
GIZA++	Och & Ney (2003)	no	no
LIU	Holmqvist & Ahrenberg (2011)	no	no
ISI2	Fraser & Marcu (2005)	no	no
JHU	Schafer & Drábek (2005)	no	no
UMIACS1	Lopez & Resnik (2005)	no	no
XRCE	Dejean et al. (2003)	no	no
1+H+F+T	This work (Section 4.1)	no	one
ProAlign	Lin & Cherry (2003a)	no	one
UMIACS2	Lopez & Resnik (2005)	no	one
1+H+F+P	This work (Section 3.2.1)	no	both
RACAI	Tufis et al. (2005)	no	both
USheffield	Aswani & Gaizauskas (2005)	no	both
ISI5	Fraser & Marcu (2005)	yes	no
Vigne	Liu et al. (2010)	yes	no

GIZA++ experiments is $1^3h^5 3^3 4^{10}$, that is, three iterations of Model 1, followed by five iterations of the HMM model, three iterations of Model 3, and ten iterations of Model 4. This is more than what Och & Ney (2003) used, again with the intention of obtaining a strong baseline. The exception is the English-Swedish data, where the HMM model turned out to produce better results, using the following configuration: 1^3h^{10}.

3.2.6. Results

To illustrate the convergence properties of the models, the experiments in this chapter were performed with an excessive number of iterations: in the order of $10^8/|S|$, where $|S|$ is the number of parallel sentences in a given corpus. In a practical setting with limited time and computational resources available, one would use considerably fewer iterations.

Almost all experiments used the same pipelined approach, where the lexical-only model (1) was initialized with alignments sampled from a uniform distribution, and the last sample of alignments from this model was used for the combined lexical and word order model (1+H), whose final sample in turn was used to initialize the model that also included fertility parameters (1+H+F). An equal number of samples was produced during each step of the process. The only exception was the English-French experiment, where

the fertility model does not lead to any improvements, so the last step of the pipeline is omitted.

The baseline system (1+H+F, or 1+H for English-French) performed very strong compared to previous results. For English-French (Table 3.3), English-Inuktitut (Table 3.5) and English-Hindi (Table 3.6), this system outperformed any system using a comparable amount of resources, and for Romanian-English (Table 3.4) and English-Swedish (Table 3.7) it was close. In the case of English-Inuktitut, the baseline system even obtained the best published results.

This is reassuring, since many of the other systems used complex models, often fine-tuned for a particular language pair, whereas my baseline system used the same parameters for all language pairs. On the other hand, it is worrying that the model did not improve upon the GIZA++ baseline for the English-Swedish alignment task. This is likely due to simplifications in the model, primarily the exclusion of NULL words. It seems that the strength of the Gibbs sampling algorithm used lies not in aligning long bitexts with similar languages, but rather the more difficult case of shorter texts in different languages.

Note that the largest data sparsity issues are found in the English-Inuktitut and English-Hindi corpora, due to the polysynthetic nature of Inuktitut and the short English-Hindi bitext, respectively. This should give the Bayesian version an advantage, and it is indeed in these two cases that we see the largest gains over GIZA++ and other non-Bayesian results.

Table 3.3.: English-French results (WPT-03 test set). $|S| = 4038$, $|P| = 17438$. 1,130,588 sentences.

Model	$\|A\|$	$\|A \cap S\|$	$\|A \cap P\|$	P	R	F	AER
Baseline system, single runs							
1	4028 ± 5.2	2853 ± 3.1	3725 ± 4.4	92.5 ± 0.1	70.7 ± 0.1	80.1 ± 0.1	18.4 ± 0.1
1+H	5312 ± 11.6	3689 ± 5.3	5079 ± 9.3	95.6 ± 0.1	91.4 ± 0.1	93.4 ± 0.1	6.2 ± 0.1
Baseline system, averaged marginals							
1	4036	2862	3739	92.6	70.9	80.3	18.2
1+H	5359	3717	5134	95.8	92.1	93.9	5.8
Best previous results (comparable)							
GIZA++	4831	3531	4715	97.6	87.4	92.2	7.0
XRCE				90.1	93.8	91.9	8.5
Best previous results (allowing additional resources)							
ProAlign				91.5	96.5	93.9	5.7
Vigne				–	–	–	4.0

Table 3.4.: Romanian-English results (WPT-05 test set). $|S| = |P| = 5034$. 48,641 sentences.

Model	$\|A\|$	$\|A \cap S\|$	$\|A \cap P\|$	P	R	F	AER
Baseline system, single runs							
1	2217 ± 9.8	1980 ± 9.4	1980 ± 9.4	89.3 ± 0.3	39.3 ± 0.2	54.6 ± 0.2	45.4 ± 0.2
1+H	3180 ± 7.9	2908 ± 10.3	2908 ± 10.3	91.4 ± 0.2	57.8 ± 0.2	70.8 ± 0.2	29.2 ± 0.2
1+H+F	3333 ± 11.4	3012 ± 9.5	3012 ± 9.5	90.4 ± 0.2	59.8 ± 0.2	72.0 ± 0.1	28.0 ± 0.1
Baseline system, averaged marginals							
1	2233	2003	2003	89.7	39.8	55.1	44.9
1+H	3222	2959	2959	91.8	58.8	71.7	28.3
1+H+F	3374	3070	3070	91.0	61.0	73.0	27.0
Best previous results (comparable)							
GIZA++	3730	3161	3161	84.7	62.8	72.1	27.9
ISI2				87.9	63.1	73.5	26.6
Best previous results (allowing additional resources)							
RACAI				76.8	71.2	73.9	26.1
Vigne				–	–	–	24.7

Table 3.5.: English-Inuktitut results (WPT-05 test set). $|S| = 293$, $|P| = 1972$. 333,185 sentences.

Model	$\lvert A\rvert$	$\lvert A \cap S\rvert$	$\lvert A \cap P\rvert$	P	R	F	AER
Baseline system, single runs							
1	410 ± 3.7	142 ± 2.7	329 ± 3.9	80.2 ± 1.1	48.5 ± 0.9	60.5 ± 0.9	33.0 ± 0.8
1+H	539 ± 12.5	245 ± 5.2	484 ± 11.4	89.8 ± 1.2	83.7 ± 1.8	86.6 ± 1.3	12.4 ± 1.2
1+H+F	560 ± 9.1	253 ± 5.2	515 ± 9.6	91.9 ± 1.1	86.5 ± 1.8	89.1 ± 1.3	9.9 ± 1.2
Baseline system, averaged marginals							
1	374	148	336	89.8	50.5	64.7	27.4
1+H	556	255	521	93.7	87.0	90.2	8.6
1+H+F	598	267	559	93.5	91.1	92.3	7.3
Best previous results (comparable)							
GIZA++	342	170	306	89.5	58.0	70.4	25.0
JHU				96.7	76.8	85.6	9.5
JHU				84.4	92.2	88.1	14.3
Best previous results (allowing additional resources)							
Vigne				–	–	–	8.9

Table 3.6.: English-Hindi results (WPT-05 test set). $|S| = |P| = 1409$. 3,556 sentences.

Model	$\|A\|$	$\|A \cap S\|$	$\|A \cap P\|$	P	R	F	AER
Baseline system, single runs							
1	589 ± 10.2	454 ± 9.2	454 ± 9.2	77.2 ± 0.8	32.3 ± 0.7	45.5 ± 0.7	54.5 ± 0.7
1+H	696 ± 11.4	575 ± 10.1	575 ± 10.1	82.5 ± 1.5	40.8 ± 0.7	54.6 ± 0.9	45.4 ± 0.9
1+H+F	730 ± 9.2	598 ± 6.4	598 ± 6.4	81.9 ± 1.2	42.5 ± 0.5	55.9 ± 0.6	44.1 ± 0.6
Baseline system, averaged marginals							
1	585	486	486	83.1	34.5	48.7	51.3
1+H	705	598	598	84.8	42.4	56.6	43.4
1+H+F	755	631	631	83.6	44.8	58.3	41.7
Best previous results (comparable)							
GIZA++	984	615	615	62.5	43.6	51.4	48.6
UMIACS1				42.9	56.0	48.6	51.4
Best previous results (allowing additional resources)							
UMIACS2				43.7	56.1	49.1	50.9
Vigne				–	–	–	44.8
USheffield				77.0	60.7	67.9	32.1

Table 3.7.: English-Swedish results (Europarl 700k sentences). $|S| = 3340$, $|P| = 4577$. 692,662 sentences.

Model	$\|A\|$	$\|A \cap S\|$	$\|A \cap P\|$	P	R	F	AER
Baseline system, single runs							
1	2079 ± 11.9	1836 ± 7.9	1887 ± 9.0	90.8 ± 0.4	55.0 ± 0.2	68.5 ± 0.3	31.3 ± 0.3
1+H	2941 ± 16.6	2581 ± 13.1	2744 ± 15.6	93.3 ± 0.4	77.3 ± 0.4	84.6 ± 0.3	15.2 ± 0.3
1+H+F	3094 ± 15.1	2663 ± 14.3	2844 ± 13.6	91.9 ± 0.3	79.7 ± 0.4	85.4 ± 0.3	14.4 ± 0.3
Baseline system, averaged marginals							
1	2110	1887	1938	91.8	56.5	70.0	29.8
1+H	3005	2645	2816	93.7	79.2	85.8	13.9
1+H+F	3183	2742	2933	92.1	82.1	86.8	13.0
Best previous results (comparable)							
GIZA++	3436	2890	3136	91.3	86.5	88.8	11.1
LIU				85.3	–	–	12.6

In order to gain some insights into the number of iterations required to obtain a given level of accuracy, Figures 3.1, 3.2, 3.3 and 3.4 show how the AER varies over long runs in different languages. The eight dotted lines in each graph represent independent runs, while the solid line uses the averaged alignment probability marginals from those eight runs.

First of all, it is clear from inspection that all through the training process, using the averaged marginals is superior to even the best single run, and even more so compared to the average single run. All of these curves are guaranteed to converge in the limit, so the same improvement could also be obtained by running a single sampler for a much larger number of iterations. However, these results show that averaging multiple concurrent samplers in this way is a more computationally efficient option. The greatest improvement is seen for the English-Hindi and English-Inuktitut corpora, which also happen to be the corpora that suffer most from data sparsity. This results in a larger proportion of word types, which, by chance, will be initialized with incorrect links that are however consistent across several different sentences, and so very unlikely to change during sampling given the sparse priors used. Unfortunately, time constraints made it impossible to explore the effects of averaging in depth.

Second, we see that performance increased only very slowly after the first few iterations, depending on the size of the training data. At the expense of some accuracy, the learning process can be shortened considerably to be competitive with EM-based algorithms in terms of computational efficiency.

3.3. Collapsed and explicit Gibbs sampling

Recall from Section 2.5.3 that in a *collapsed* Gibbs sampler, some variables are marginalized out. In the word alignment models described so far, the marginalized variables are the categorical distribution parameters, so that the only variables sampled are the alignment links.

An explicit sampler would have to sample these variables *and* the various categorical distributions, of which the largest are the lexical translation distributions $p_t(f|e)$ where the number of parameters is the product of the vocabulary sizes of the two languages. For relatively small corpora, sampling the categorical distributions therefore dominates the computational work during learning.

On the other hand, in a collapsed Gibbs sampler each alignment variable is conditioned on all other alignment variables, including those previously sampled. This dependency makes it difficult to parallelize the alignment process, which is becoming an increasingly serious downside as the level of parallelism grows in modern computer architectures. An explicit sampler, in contrast, can sample the categorical distributions in parallel, then sample alignment variables for all sentences in parallel. Zhao & Gildea (2010) utilize this parallelism in their word alignment algorithm where the expectation step of the EM algorithm is computed using Gibbs sampling.

If explicit Gibbs sampling produces competitive results, it would be a very attractive option. Unfortunately, to my knowledge there have been no published studies exploring

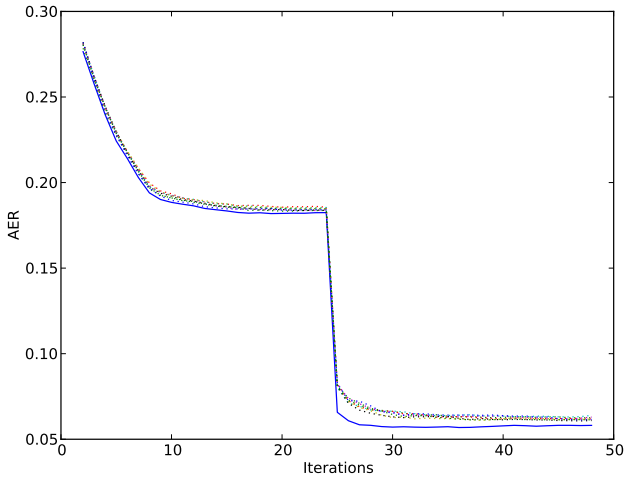


Figure 3.1.: AER during training (English-French), see Table 3.3.

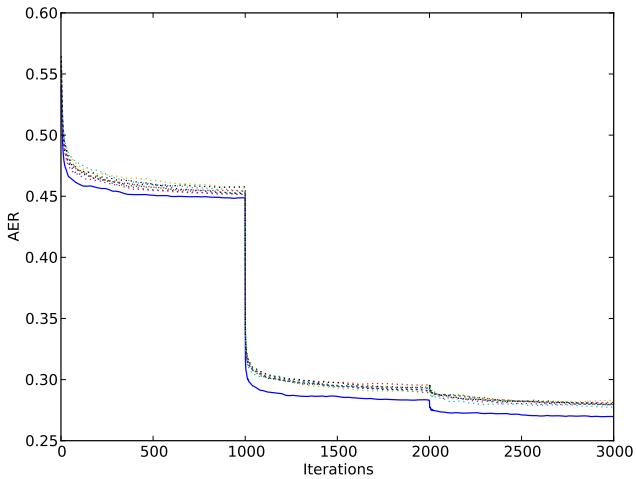


Figure 3.2.: AER during training (Romanian-English), see Table 3.4.

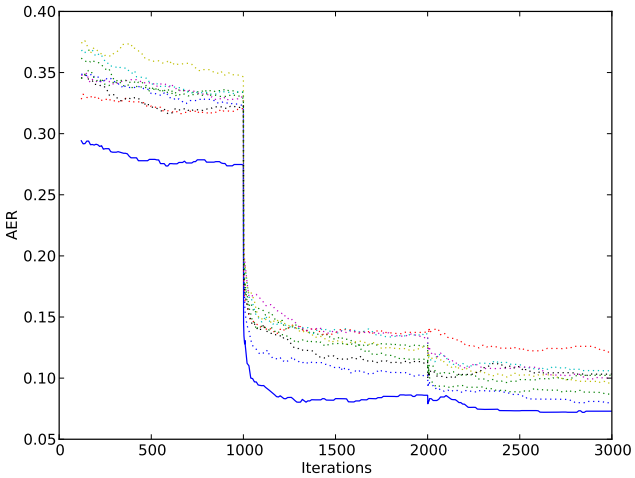


Figure 3.3.: AER during training (English-Inuktitut), see Table 3.5.

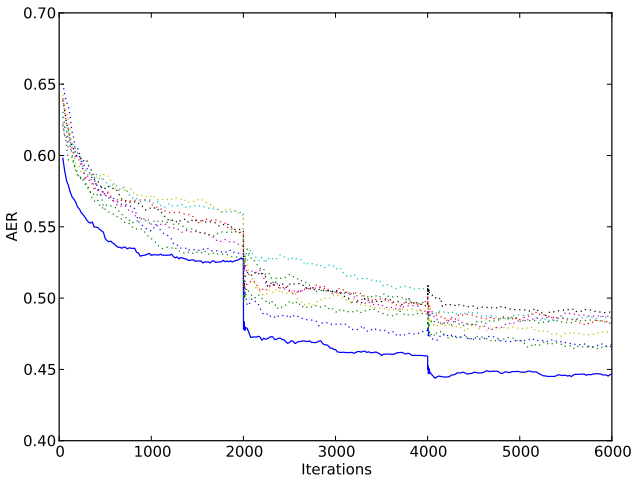


Figure 3.4.: AER during training (English-Hindi), see Table 3.6.

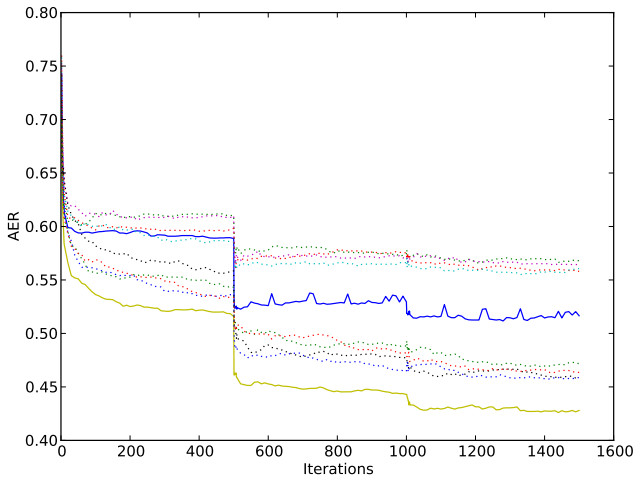


Figure 3.5.: AER during training (English-Hindi), with collapsed sampling (bottom) and explicit sampling (top).

collapsed versus explicit Gibbs sampling algorithms for the problem of word alignment.

3.3.1. Experiments

In order to test the feasability of explicit Gibbs sampling, I implemented an explicit version of the algorithms from Section 3.2.1. Figure 3.5 shows the result of both variants of the alignment algorithm on identical data. Since the categorical distribution sampling required a time proportional to the vocabulary sizes of the two languages multiplied, I chose to only perform this evaluation on a relatively short text (English-Hindi, see Table 3.1) with a vocabulary of manageable size.

The same number of iterations (500 per model) and the same chain of models (1, 1+H, 1+H+F) were used in both cases, and it is clear from the figure that the collapsed sampler performed much better. The explicit sampler quickly reached a certain level of accuracy and then made no perceivable progress during the learning process. In contrast, the collapsed sampler gradually improved over hundreds of iterations. To explain this result, we can look at the leftmost part of Figure 3.5, where only lexical translation distributions are used. Consider an example where the only instance of the source word *dog* is incorrectly aligned to the target word *Katze* 'cat.' In a collapsed sampler, the alignment link for *Katze* is sampled conditioned on all other alignment links, *excluding*

the present, incorrect link. This means that the probability of aligning to *dog* is about the same as the correct alignment to *cat*, which is as expected given that co-occurrence is the only clue available to the algorithm.

On the other hand, the explicit sampler first samples a categorical distribution for the translation of *dog* based on the incorrect alignment link, which (given a sparse Dirichlet prior) is extremely likely to result in a categorical distribution where *Katze* has a much higher probability than any other word, including the correct option in this case: *Hund*. When the alignment variable is then sampled, the incorrect alignment to *Katze* is chosen again with very high probability.

This resistance to moving between similar and equally probable solutions causes very slow mixing, and even 1,500 iterations is not enough to find a noticeably better solution. Note that the theoretical guarantees of the MCMC algorithm are met in both cases: as the number of samples grows towards infinity, *Katze* aligns to *dog* and *cat* about the same number of times. The difference is only that the collapsed sampler reaches this distribution faster.

3.4. Alignment pipeline

Previous research with EM-based alignment models has shown that initialization through a "pipeline" of successively more complex models is essential to good performance (Och & Ney 2003, pp 36–37). Gal & Blunsom (2013) confirmed that this also holds for their Bayesian versions of the IBM models using Gibbs sampling, but only for a single corpus and only comparing two different pipelines.

In this section, I evaluate three configurations: a fully pipelined training (abbreviated 1+H+F), model 1 followed by HMM/fertility (abbreviated 1+HF), and the most complex model with random initialization (abbreviated 1HF). Given an equal number of iterations, pipelined training is actually faster since some of the time is spent with simpler and computationally less demanding models. Therefore, as long as pipelined training does not significantly decrease accuracy, we should prefer it. Figures 3.6 and 3.7 show that this is not the case, but rather that AER somewhat improves as the pipeline gets longer. In spite of the very different characteristics in terms of size and language similarity between these two datasets, the training curves are strikingly similar. Based on these results, I continued to use fully pipelined training for the remainder of the experiments in this thesis.

3.5. Choice of priors

Several of the IBM alignment models contain a large number of parameters, which make the models prone to overfitting and sensitive to data sparsity. Various ways have been tried to alleviate this problem, such as more compact reparameterizations (Dyer et al. 2013) or ad-hoc smoothing schemes to combat data sparsity (Och & Ney 2003, section 6.5).

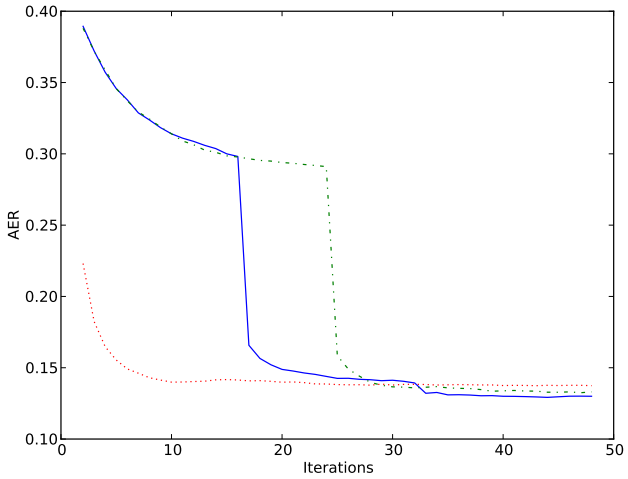


Figure 3.6.: AER during training (English-Swedish), using pipelines 1HF (dotted), 1+HF (half-filled) and 1+H+F (filled).

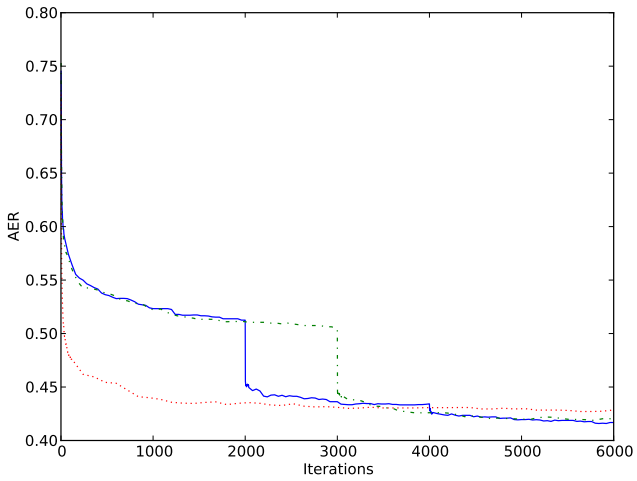


Figure 3.7.: AER during training (English-Hindi), using pipelines 1HF (dotted), 1+HF (half-filled) and 1+H+F (filled).

A more systematic approach is to use a Bayesian model with hierarchical priors (see Section 2.4.6), as was done by Gal & Blunsom (2013). For instance, if we use a word order model $p_j(l|c(w_{j-1}))$ that depends on the jump length $l = a_j - a_{j-1}$ as well as the word class $c(f_j)$ of word f_j, we can assume a prior with $f_j(l)$ as its back-off distribution:

$$p_j(l|c(w_{j-1})) \sim \text{PYP}(d, \alpha, p_j(l))$$

where $p_j(l)$ only depends on the jump length l.

Yarin Gal (p.c.) recalls that there was some improvement when going from a non-hierarchical to a hierarchical model for the Bayesian HMM model of Gal & Blunsom (2013), but no figures have been published. In this section, I perform empirical evaluations comparing corresponding hierarchical and non-hierarchical models.

3.5.1. Algorithms

There are many possible ways to turn a distribution conditioned on a number of variables into a hierarchical model. I generally follow the hierarchical structures of Gal & Blunsom (2013), but since our respective basic probability models differ, the hierarchical models are also not equal.

The lexical distribution for each target word f depends on the source word e, with a backoff to the target word distribution, and then to the uniform distribution U:

$$p_t(f|e) \sim \text{PYP}(d_f^t, \alpha_f^t, p_t(f))$$
$$p_t(f) \sim \text{PYP}(d_0^t, \alpha_0^t, U)$$

For the word order model, I assume the jump length $l = a_j - a_{j-1}$ depends on the word class bigram $c(w_{j-1}), c(w_j)$ and the sentence length I with a backoff to the class unigram $c(w_j)$ and length, then to the sentence length only, and finally to the uniform distribution:

$$p_j(l|I, c(w_{j-1}) \sim \text{PYP}(d_{Ic}^j, \alpha_{Ic}^j, p_j(l|I))$$
$$p_j(l|I) \sim \text{PYP}(d_I^j, \alpha_I^j, p_j(l))$$
$$p_j(l) \sim \text{PYP}(d_0^j, \alpha_0^j, U)$$

As for fertility parameters, there is considerable sparsity since separate fertility counts are used for each word type. By using a common feritilty distribution as a prior for individual word fertilities, we should be able to capture the general level of synthesis in the source language:

$$p_f(\phi|e) \sim \text{PYP}(d_e^f, \alpha_e^f, p_f(\phi))$$
$$p_f(\phi) \sim \text{PYP}(d_0^f, \alpha_0^f, U)$$

Finally, the hierarchical tag translation distribution is constructed in the same way as the lexical translation distribution:

$$p_p(t_f|t_e) \sim \text{PYP}(d_t^p, \alpha_t^p, p_p(t_f))$$
$$p_p(t_f) \sim \text{PYP}(d_0^p, \alpha_0^p, U)$$

This model introduces a number of new parameters: the discount and concentration parameters, d and α, for each of the Pitman-Yor priors. These are sampled using slice sampling (see Section 2.5.6), with a uniform prior for the discount d and an exponential prior $p(\alpha) = \lambda e^{-\lambda x}$ for the concentration parameters. This, of course, introduces yet another set of new parameters—but the model performance is not very sensitive to the values of λ, and so they are all set to 1 except the prior for α_f^t, where it is necessary to enforce sparsity by setting this prior to 100.

3.5.2. Evaluation setup

To test whether the models with hierarchical Pitman-Yor priors perform better than those with the corresponding non-hierarchical Dirichlet priors, I evaluated each under identical conditions. These differed from the evaluations in Section 3.2 in two ways. First, the number of iterations was reduced, since more iterations resulted either in no improvement or even a reduction of accuracy for the hierarchical models. Second, the soft growing symmetrization (Equation (3.3)) optimized on the respective development sets was used, rather than the soft intersection symmetrization (Equation (3.2)). This is in line with most of the baselines, which also report the best result among multiple symmetrization methods, and helps to ensure fair comparisons between different types of algorithms.

For the two language pairs for which I had access to supervised PoS taggers, English-French and English-Swedish, I also performed the corresponding experiment using the output of these taggers. The tags were used both in the HMM model (H) to condition the jump probabilities and in the tag translation model (P). In addition, lemmas (Swedish) or stems (English) were used instead of the full word forms, further improving accuracy.

3.5.3. Results

Table 3.8 shows that the hierarchical models (with L superscripts) outperformed the corresponding non-hierarchical models across the different corpora. Compared to the previously published results on these data sets (summarized in Section 3.2.6), we can see that the hierarchical models were quite accurate. In all cases except English-Swedish, they achieved the lowest AER among unsupervised systems, and in several cases they were on par with or better than semi-supervised systems or systems using language-specific resources.

The main exception to this pattern is that the English-Hindi aligner of Aswani & Gaizauskas (2005) performed better, because they had access to resources such as an English-Hindi bilingual dictionary, a gazetteer of Hindi named entities, and a transliteration tool. This is particularly important since the English-Hindi parallel text is very short, only 3,556 sentences, which makes unsupervised learning very challenging. The Bayesian models were still much more accurate than the unsupervised baselines, and even the semi-supervised Vigne system (Liu et al. 2010).

Due to the increased complexity of the hierarchical Pitman-Yor priors, alignment was roughly three times slower than when using the non-hierarchical Dirichlet priors. This

tradeoff is important to keep in mind, since in some circumstances it would be better to use simple Dirichlet priors with an increased number of iterations and/or increased number of independent samplers that can be averaged.

Table 3.8.: Hierarchical Pitman-Yor priors (with L superscript) compared to the corresponding non-hierarchical Dirichlet priors (without superscript). Refer to Section 3.2.6 for relevant baselines.

Model	$\|A\|$	$\|A \cap S\|$	$\|A \cap P\|$	P	R	F	AER
English-French (5 iterations/model)							
1+H	5361	3725	5086	94.9	92.2	93.5	6.3
1^{L}+H^{L}	5689	3775	5378	94.5	93.5	94.0	5.9
1+H+F+P	5830	3810	5530	94.9	94.4	94.6	5.4
1^{L}+H^{L}+F^{L}+P^{L}	5675	3744	5471	96.4	92.7	94.5	5.1
Romanian-English (25 iterations/model)							
1+H+F	3945	3274	3274	83.0	65.0	72.9	27.1
1^{L}+H^{L}	4303	3509	3509	81.5	69.7	75.2	24.8
English-Inuktitut (8 iterations/model)							
1+H+F	446	237	428	96.0	80.9	87.8	10.0
1^{L}+H^{L}+F^{L}	532	257	511	96.1	87.7	91.7	6.9
English-Hindi (50 iterations/model)							
1+H+F	1023	687	687	67.2	48.8	56.5	43.5
1^{L}+H^{L}	931	684	684	73.5	48.5	58.5	41.5
English-Swedish (5 iterations/model)							
1+H+F	3679	2874	3160	85.9	86.0	86.0	14.0
1^{L}+H^{L}+F^{L}	3573	2882	3182	89.1	86.3	87.7	12.3
1+H+F+P	3529	2938	3201	90.7	88.0	89.3	10.6
1^{L}+H^{L}+F^{L}+P^{L}	3457	2938	3207	92.8	88.0	90.3	9.6

4. Word alignment and annotation transfer

4.1. Aligning with parts of speech

Several authors, starting from Brown et al. (1993), have used word classes to aid word alignment in various manners. Toutanova et al. (2002) showed that if both parts of a bitext are annotated with PoS tags, alignment accuracy can be improved simply by using a tag translation model $p(t_f|t_e)$ in addition to the word translation model $p(f|e)$.

Good PoS annotations for both languages may be a realistic scenario when two well-resourced languages with good supervised PoS taggers (or sufficient unannotated data for unsupervised PoS taggers) are to be aligned, but not so much for the vast majority of human languages. However, Yarowsky & Ngai (2001) showed that by selecting reliable alignments, it is possible to accurately transfer PoS tag annotations from one language to the other in a bitext. By simultaneously learning both alignments and target language PoS tags, it should therefore be possible to benefit both tasks.

4.1.1. Naive model

The most direct approach to integrating PoS transfer into the Bayesian alignment method described in Section 3.2.1 is to include a tag translation distribution $p_c(t_f|t_e)$, following Toutanova et al. (2002), and simultaneously sampling alignment variable a_j and the corresponding tag t_j of each target word j.

There is however one crucial difference between the work of Toutanova et al. (2002) and the present work: while they assume both source and target language tags to be fixed, the point here is to learn the target-side tags. It is easy to see that (all other things equal) the model just outlined would assign highest probability to solutions where all target words have the same tag—which is clearly undesirable!

4.1.2. Circular generation

The naive model proposed above describes a situation where a source word generates a target word, and the corresponding source word tag independently generates a tag for the target word.

I have investigated a model where words and tags are instead generated in a "circle" (see Figure 4.1), starting with the source word generating a target word, which generates a target tag, which finally generates the source tag. Of these four variables, only the target tag is unobserved.

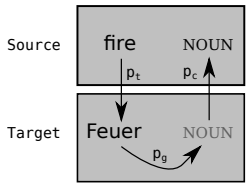


Figure 4.1.: Conditional dependencies in the circular generation model. All variables except the target-side PoS tag (in gray) are observed. Non-lexical dependencies are left out for clarity.

$$P(\boldsymbol{f}, \boldsymbol{a}, \boldsymbol{t}_e, \boldsymbol{t}_f | \boldsymbol{e}) \propto \prod_{j=1}^{J} p_t(f_j | e_{a_j}) p_g(t_{f_j} | f_j) p_c(t_{e_{a_j}} | t_{f_j}) \tag{4.1}$$

Equation (4.1) gives the total probability of an alignment under the most basic circular model, which does not include word order or fertility. These are however easy to include as separate factors (like IBM Model 1 in Equation (2.2) was extended to the HMM model in Equation (2.4)), since they depend only on the alignment variables $\boldsymbol{a}$ and not on the lexical $\boldsymbol{f}$ or PoS tag $\boldsymbol{t}_f$ variables.

Note that the final term, $p_c(t_{e_{a_j}} | t_{f_j})$, depends on both the alignment variable a_j and the target-side tag t_{f_j}. While this coupling comes at the cost of sampling each a_j, t_{f_j} pair requiring $\mathrm{O}\big(T \cdot |\boldsymbol{e}|\big)$ rather than $\mathrm{O}\big(T + |\boldsymbol{e}|\big)$ operations, the goal is increased consistency between tags and alignments, hopefully to the benefit of both.

4.1.3. Alternating alignment-annotation

Yarowsky & Ngai (2001) developed a method for annotation transfer, where PoS tags were projected through word alignments (both assumed to be given), and selected subsets of the projected tags were used to estimate, using different heuristics, the emission and transition probabilities of a HMM tagging model for the target language.

They concluded that while projected tags were too noisy for training accurate tagging models, very good models could be obtained by using tags from sentences with high-probability alignments. Tag transition distributions in particular contain few parameters and can be estimated reliably from small amounts of data, which allows a considerable amount of recall to be traded for high precision in the projection step.

The simplest method to combine the PoS annotation transfer algorithm of Yarowsky & Ngai (2001) with a PoS-aware alignment algorithm in the style of Toutanova et al. (2002) would be to just perform them alternatively.

It is also possible to extend the Gibbs sampling alignment algorithms with ideas from both of these works, resulting in the following scheme for sampling alignment and PoS tag variables:

1. Sample all alignments, given the current target PoS tags.
2. Estimate transition and emission parameters of a HMM given the current alignments.
3. Sample target PoS tags from the HMM.

From these samples, any of the marginalization methods from Section 2.5.5 may be used to decide on the final alignment and/or PoS tags, whichever is desired.

Since the basic alignment algorithm is asymmetric, both alignment directions are sampled in parallel in step 1 above. After this, the last such sample is symmetrized as in Equation (3.2), and the resulting links are used to estimate the transition parameters from the 50% most reliably aligned sentences, and the emission parameters from all sentences. I use the anti-smoothing heuristic of Yarowsky & Ngai (2001) as well as the affix-tree algorithm of Schmid (1994) to exploit morphological information. To capture the fact that there are both suffixing and prefixing languages, both suffix trees and prefix trees are tried, and the one resulting in the largest information gain on the training data is used for the final tagging.

Annotation transfer for resource-rich languages has evolved considerably since the pioneering work of Yarowsky & Ngai (2001), and recent work exploits a variety of resources in addition to parallel texts, e.g. crowd-sourced dictionaries (Täckström et al. 2013) or label propagation with large monolingual corpora (Das & Petrov 2011) to improve accuracy. Although some of these authors claim that their goal is to transfer annotations to resource-poor languages, these languages may be "poor" only compared to English. The most typical case among the 7,000 or so languages of the world is that there are no electronic parallel texts, dictionaries, or even monolingual texts of reasonable size. For the 1,000+ languages where electronic New Testament translations are available, this is typically the *only* such resource useful in annotation transfer.

For this reason, I focus on methods that require only a short parallel text like the New Testament. These methods will naturally not be competitive with specialized algorithms for languages in the medium-to-rich resource range. The same applies to the method I use for PoS tagging: whereas more advanced algorithms have been developed in the decades since Schmid (1994), increased accuracy when learning from a high-quality annotation does not necessarily translate into increased accuracy when learning from a small and noisy set of transferred PoS tag annotations.

4.2. Evaluation

There are three ways of evaluating the alternating alignment-annotation method. First, one could annotate the target language text with gold standard PoS tags and compare these to the result of the annotation transfer. Unfortunately, I do not have access to any such data (except for the study described in Section 4.3), but given the good performance of supervised PoS taggers it is possible to automatically annotate the target language text where a tagger is available, at the expense of some additional uncertainty.

The second option is to use the model estimated from the parallel text to tag a monolingual text that has gold standard PoS annotations. Since I work with the Universal PoS Tags of Petrov et al. (2012), the Universal Dependency Treebank (McDonald et al. 2013) provides a suitable evaluation set.

Note that the first method estimates in-domain performance, while the second method could also estimate out-of-domain performance if the genres differ between the parallel text and the evaluation data.

Finally, we can also evaluate the general success of the PoS transfer by investigating the change in alignment quality after the resulting PoS tags were introduced into the alignment model (see Section 3.2.1.1). In applications where the PoS tags are not needed, this is in fact the most important measure.

4.2.1. Alignment quality evaluation

In this evaluation, there is one natural baseline, the 1+H+F algorithm of Section 3.2.1 using lexical, word order and fertility variables. There is also one natural upper bound,[1] 1+H+F+P, which additionally includes high-quality PoS tags from supervised taggers for both languages. Such taggers are however only available for both languages in the English-French and English-Swedish tasks, so the remaining data sets unfortunately contain no such figures. The alternating alignment-annotation method, 1+H+F+T, is expected to score somewhere between the 1+H+F and 1+H+F+P models. In addition, relevant results from previous work (repeated from Section 3.2) are also given.

In terms of alignment quality, it is clear that annotation transfer is beneficial for all the language pairs evaluated: English-French (Table 4.1), Romanian-English (Table 4.2), English-Inuktitut (Table 4.3), English-Hindi (Table 4.4) as well as English-Swedish (Table 4.5). Please refer back to Section 3.2 for details on how these tables are to be interpreted.

In all these evaluations the alternating alignment-annotation algorithm is comparable to or better than the best systems using equivalent resources. This is very promising, since many of those systems use a variety of dictionaries and other resources, whereas the alignment-annotation system only has access to a PoS tagger for one of the languages (English, in all these cases). The largest improvement can be seen in the English-Hindi task, which due to the small size of the data and the fairly large difference between the languages is the most difficult.

4.2.2. Annotation quality evaluation

While maximizing the accuracy of the transferred PoS tags is not my primary goal, it is still useful to evaluate the level of performance achieved with the alternating alignment-annotation. First of all, we might want to know what level of PoS tagging accuracy is needed to give the alignment performance gains observed in Section 4.2.1. Second, it

[1] Technically, this is not a proper "upper bound" since it could be exceeded in theory, for instance, if the joint PoS transfer and alignment algorithm produces tags that are in fact *better* than the actual PoS tags of the target language for the purpose of guiding word alignment.

Table 4.1.: English-French results (WPT-03 test set). $|S| = 4038$, $|P| = 17438$. 1,130,588 sentences.

Model	$\|A\|$	$\|A \cap S\|$	$\|A \cap P\|$	P	R	F	AER
Baseline (1+H) and alternating alignment-annotation (1+H+T) and using supervised PoS tagging (1+H+P).							
1+H	5359	3717	5134	95.8	92.1	93.9	5.8
1+H+T	5505	3751	5254	95.4	92.9	94.1	5.6
1+H+P	5542	3778	5263	95.0	93.6	94.3	5.6
Best previous results (comparable or fewer resources)							
GIZA++	4831	3531	4715	97.6	87.4	92.2	7.0
XRCE				90.1	93.8	91.9	8.5
Best previous results (allowing additional resources)							
ProAlign				91.5	96.5	93.9	5.7
Vigne				–	–	–	4.0

Table 4.2.: Romanian-English results (WPT-05 test set). $|S| = |P| = 6201$. 48,641 sentences.

Model	$\|A\|$	$\|A \cap S\|$	$\|A \cap P\|$	P	R	F	AER
Baseline (1+H+F) and alternating alignment-annotation (1+H+F+T)							
1+H+F	3374	3070	3070	91.0	61.0	73.0	27.0
1+H+F+T	3447	3120	3120	90.5	62.0	73.6	26.4
Best previous results (comparable or fewer resources)							
GIZA++	3730	3161	3161	84.7	62.8	72.1	27.9
ISI2				87.9	63.1	73.5	26.6
RACAI				76.8	71.2	73.9	26.1
Best previous results (allowing additional resources)							
Vigne				–	–	–	24.7

Table 4.3.: English-Inuktitut results (WPT-05 test set). $|S| = 293$, $|P| = 1972$. 333,185 sentences.

Model	$\|A\|$	$\|A \cap S\|$	$\|A \cap P\|$	P	R	F	AER
Baseline (1+H+F) and alternating alignment-annotation (1+H+F+T)							
1+H+F	598	267	559	93.5	91.1	92.3	7.3
1+H+F+T	630	273	595	94.4	93.2	93.8	6.0
Best previous results (comparable or fewer resources)							
GIZA++	342	170	306	89.5	58.0	70.4	25.0
JHU				96.7	76.8	85.6	9.5
JHU				84.4	92.2	88.1	14.3
Best previous results (allowing additional resources)							
Vigne				–	–	–	8.9

Table 4.4.: English-Hindi results (WPT-05 test set). $|S| = |P| = 1409$. 3,556 sentences.

Model	$\|A\|$	$\|A \cap S\|$	$\|A \cap P\|$	P	R	F	AER
Baseline (1+H+F) and alternating alignment-annotation (1+H+F+T)							
1+H+F	712	606	606	85.1	43.0	57.1	42.9
1+H+F+T	817	677	677	82.9	48.0	60.8	39.2
Best previous results (comparable or fewer resources)							
GIZA++	984	615	615	62.5	43.6	51.4	48.6
UMIACS2				43.7	56.1	49.1	50.9
Best previous results (allowing additional resources)							
Vigne				–	–	–	44.8
USheffield				77.0	60.7	67.9	32.1

Table 4.5.: English-Swedish results (Europarl 700k sentences). $|S| = 3340$, $|P| = 4577$. 692,662 sentences.

Model	$\|A\|$	$\|A \cap S\|$	$\|A \cap P\|$	P	R	F	AER
Baseline (1+H+F), alternating alignment-annotation (1+H+F+T), and using supervised PoS tagging (1+H+F+P).							
1+H+F	3183	2742	2933	92.1	82.1	86.8	13.0
1+H+F+T	3125	2774	2961	94.8	83.1	88.5	11.3
1+H+F+P	3262	2823	3034	93.0	84.5	88.6	11.3
Best previous results (comparable or fewer resources)							
GIZA++	3436	2890	3136	91.3	86.5	88.8	11.1
LIU				85.3	–	–	12.6

is useful to know what kind of accuracy can be expected when transfering annotations to severely under-resourced languages, where the only resource available might be a translation of the New Testament.

Table 4.6 shows the accuracy of PoS transfer with two rather different types of corpora: large collections of parliamentary proceedings (roughly a million sentences each), and pairs of translations from the New Testament corpus, which are roughly 8,000 verses each. The out-of-domain figures are lower than current state of the art for unlimited-resources PoS transfer for the given languages: Täckström et al. (2013) reported 88.3% and 88.9% accuracy for English-French and English-Swedish respectively (but note that due to very different evaluation setups, their figures are not directly comparable to Table 4.6). What the results do show, however, is that jointly learning PoS tags and alignments not only benefits alignment quality (as was shown in Section 4.2.1), but often has a small positive effect on PoS accuracy. This tendency is not without exceptions, though, and in several cases the difference is none or very small.

One possible reason for this could be that the PoS transfer algorithm is not very sensitive to alignment quality, and the relatively modest gains in alignment quality might not be enough to significantly increase PoS transfer accuracy. It is also important to remember that a PoS tagging that is informative to the alignment process does not necessarily have to be linguistically accurate. For instance, the German verb *heissen* 'to be named' might be aligned frequently to the English noun *name*. If the tag for *heissen* is (incorrectly) sampled as NOUN, this would increase the probability of correctly aligning *heissen* and *name*, which in turn might make the PoS transfer model even more likely to mis-tag *heissen*.

Compared to the evaluation on the target side of the bitext itself, accuracy is lower when evaluated on the Universal Dependency Treebank test set (McDonald et al. 2013) of the given language, which is expected since this data is not part of the bitext and contains different genres than the bitexts. This effect is particularly large for the New Testament experiments, since the model is trained on a genre that is very different from the test data. A better morphological model in the PoS tagger could probably bridge some of this difference, but the fact remains that out-of-domain PoS tagging is a difficult problem.

Table 4.7 shows the accuracy of PoS transfer with the New Testament corpus, both using multi-source (rightmost column) and single-source (all other columns) transfer. It is clear that multi-source transfer was superior both to the median and usually even to the single best of the 22 source texts.[2] The average accuracy (bottom row) increased from the single-best result of 85.3% to the multi-source accuracy of 86.5%. The gap to the median single-source accuracy of 81.5% is even greater: multi-source transfer resulted in a 27% error reduction. The only systematic exception to this trend is when transfering tags to English, where single-source transfer from the best translation (in Swedish) gave somewhat better accuracy than multi-source transfer.

The accuracies in Table 4.7 are promising, because they show that high accuracy can

[2]Translations into the same language as the target are not used, so e.g. German "only" uses $22 - 8 = 14$ source texts.

Table 4.6.: Accuracy of PoS transfer, evaluated on the Universal Dependency Treebank test set for the target language (*UDT*) or the test set of the parallel data itself (*Test*), using tags from a supervised tagger as gold standard (when one is available). The former contains (potentially) out-of-domain text, while the latter obviously contains text from the same corpus and domain as the training data. PoS transfer using 1+H+F alignments is used as a baseline, to investigate the effect of joint alignment + PoS transfer.

	Joint		**Baseline**	
Corpus	**UDT**	**Test**	**UDT**	**Test**
English-French (WPT-03)	83.7%	85.9%	82.7%	85.8%
English-Swedish (Europarl)	84.5%	86.6%	84.2%	86.0%
English-German (NT)	73.4%	84.4%	73.4%	84.5%
English-Finnish (NT)	71.1%	–	69.0%	–
English-French (NT)	75.9%	82.4%	76.0%	82.6%
English-Indonesian (NT)	79.7%	–	76.4%	–
English-Italian (NT)	74.1%	–	73.6%	–
English-Spanish (NT)	75.6%	–	76.3%	–
English-Swedish (NT)	74.6%	86.4%	74.3%	86.3%
Swedish-English (NT)	71.8%	86.1%	70.6%	86.4%

be achieved even with the relatively short New Testament text, and in the absence of any external resources such as dictionaries. Unfortunately, I was only able to evaluate PoS tag accuracy for a small set of closely related languages, and we can assume that for the majority of the 1,001 languages in the corpus accuracy would be considerably lower if the same set of source languages was used. It would be highly desirable to include PoS taggers from a more typologically diverse set of languages in these experiments, but time constraints have made this impossible within the scope of this thesis.

Table 4.7.: PoS accuracy (in percent) using single-source (first five columns) and multi-source (rightmost column) transfer in the New Testament corpus. Rows are target texts and columns are source languages. For each language (with number of translations), the worst/median/best results are given for the different translations. The **All** columns summarize the results over all the source texts from the preceding columns. Finally, **Multi** is the result of multi-source transfer using the sums of tag marginal distributions. The best result on each row is bold-faced.

Target	**Source texts**															**Multi**
	deu (8)			**eng (5)**			**fra (5)**			**swe (4)**			**All (22)**			
deu_1				79.0	80.1	80.9	80.1	81.1	81.4	75.0	83.7	85.1	75.0	81.0	85.1	**86.8**
deu_2				79.3	80.8	81.4	79.5	80.5	80.7	78.3	83.5	85.2	78.3	80.7	85.2	**85.8**
deu_3				79.8	80.6	81.7	81.1	81.9	82.0	77.1	84.4	85.9	77.1	81.7	85.9	**88.2**
deu_4				80.6	81.1	82.4	81.0	81.8	81.8	75.3	83.2	85.4	75.3	81.6	85.4	**87.3**
deu_5				80.2	80.7	81.7	81.3	81.6	82.2	76.5	84.9	85.9	76.5	81.5	85.9	**86.8**
deu_6				79.4	81.3	81.9	80.0	80.7	81.3	80.7	85.0	**86.0**	79.4	81.0	**86.0**	85.4
deu_7				79.9	81.8	82.3	81.1	81.2	81.9	76.6	85.6	86.3	76.6	81.7	86.3	**86.4**
deu_8				80.0	81.4	82.3	81.0	81.4	82.0	76.3	84.8	85.7	76.3	81.6	85.7	**86.5**
eng_1	74.2	76.2	79.4				76.2	77.0	77.8	76.6	81.5	81.7	74.2	76.7	81.7	**83.8**
eng_2	79.2	81.8	84.2				80.9	81.4	82.0	79.8	85.8	**86.2**	79.2	81.8	**86.2**	85.5
eng_3	80.1	81.7	83.7				80.6	81.0	81.6	80.7	85.7	**86.3**	80.1	81.6	**86.3**	85.4
eng_4	79.5	81.4	84.3				80.3	80.7	81.3	78.8	85.8	**86.5**	78.8	81.3	**86.5**	85.1
eng_5	80.3	81.5	84.1				80.7	81.0	81.8	80.3	86.0	**86.7**	80.3	81.5	**86.7**	84.7
fra_1	80.2	82.9	83.5	80.1	81.3	81.5				78.0	83.8	84.5	78.0	82.5	84.5	**85.8**
fra_2	80.3	83.5	84.5	80.7	80.9	81.1				77.0	83.6	84.7	77.0	83.0	84.7	**86.0**
fra_3	80.5	83.1	83.5	79.9	81.2	81.9				77.6	84.3	85.1	77.6	82.7	85.1	**85.3**
fra_4	80.3	83.5	83.8	80.0	81.1	81.2				77.4	84.1	84.6	77.4	82.7	84.6	**85.3**
fra_5	80.5	83.2	83.8	80.1	81.2	81.4				77.2	84.0	84.8	77.2	82.9	84.8	**86.1**
swe_1	80.4	81.2	81.8	82.8	84.0	85.4	80.2	81.0	81.9				80.2	81.2	85.4	**90.3**
swe_2	76.3	77.5	79.4	80.9	81.7	82.4	76.9	77.9	79.4				76.3	78.3	82.4	**85.7**
swe_3	81.3	82.1	82.5	83.4	85.4	86.4	81.1	82.5	82.8				81.1	82.5	86.4	**90.6**
swe_4	81.8	82.3	82.7	82.7	84.8	85.4	81.8	82.7	83.3				81.8	82.7	85.4	**90.7**
Avg.	79.6	81.6	82.9	80.5	81.7	82.4	80.2	80.9	81.5	77.7	84.4	85.4	77.9	81.5	85.3	**86.5**

4.3. Tagging the Swedish Sign Language Corpus

The Swedish Sign Language Corpus (SSLC) (Mesch et al. 2014; Mesch & Wallin 2015) is a corpus of Swedish Sign Language (SSL), containing 25 hours of recorded and partially transcribed spontaneous conversation from 42 different signers. Its annotations include (among other things) a gloss for each sign and a translation into Swedish, which in effect makes it a parallel corpus of transcribed SSL and written Swedish. The version used in my experiments contains 24,976 SSL tokens, which are not sentence-segmented, and 41,910 Swedish tokens divided into 3,522 sentences.

Segmenting spontaneous SSL conversation into sentences or utterances is not a trivial task (Börstell et al. 2014), and there is currently no such segmentation in the corpus. In order to be able to use sentence-based word alignment models, I follow Sjons (2013) in using the Swedish sentences as a basis for segmentation. This is possible since both translations and glosses are associated with time slots, so that SSL glosses overlapping in time with a Swedish sentence can be segmented into a "sentence."

The SSLC originally lacked PoS tag annotations, due to both the time required for manual annotation and to theoretical problems in defining parts of speech in SSL. The research survey of Ahlgren & Bergman (2006) presented a rough division into eight parts of speech with some discussion of each: nouns, verbs, adjectives, adverbs, numerals, pronouns, conjunctions and prepositions. This classification is rather coarse, and I follow it mainly to stay in line with previous work. It also happens to be a subset of the "universal" tagset of Petrov et al. (2012), which could benefit future multilingual investigations. Lars Wallin (p.c.) suggests an extended set of classes, distinguishing for instance between simple and polysynthetic verbs.[3]

Ahlgren & Bergman's classification was also used by Sjons (2013) in his preliminary study of PoS induction in SSL, the only such study published to date. Since only monolingual unsupervised methods were used, namely Brown clusters (Brown et al. 1992) and k-means clustering (MacQueen 1967), results were predictably poor given the limited amount of data available.

In this study, I investigate whether transfer of annotation is a practical way of annotating SSLC with PoS tags, and whether jointly learning word alignments and PoS tags can improve accuracy. While the SSLC is unique as a parallel corpus of SSL and Swedish, it has a few shortcomings from the point of view of automatic word alignment and annotation transfer: the lack of sentence segmentation on the SSL side, a fairly non-literal translation into Swedish, and limited size. While the 96% accuracy reported by Yarowsky & Ngai (2001) is clearly out of reach, I have a more modest goal of reaching an accuracy high enough to make semi-automatic PoS annotation practical.

4.3.1. Data processing

The SSLC was annotated using the ELAN software (Wittenburg et al. 2006). ELAN annotations are arranged into *tiers*, each containing a sequence of annotations with time

[3]The term "polysynthetic" in this context has been largely replaced by "classifier construction", but is used here for consistency with the SSLC documentation. See e.g. Emmorey (2003) for an overview.

slots. For the present study, two types tiers are of interest: the signs of the dominant hand (which, redundantly, also includes signs by the other hand during dominance reversal), and the Swedish sentences. Signs are transcribed using glosses, whose names are most often derived from a corresponding word or expression in Swedish. Each gloss may also have a number of properties marked, such as which hand it was performed with, whether it was reduplicated, interrupted, and so on. The annotation conventions are described in further detail by Wallin et al. (2014).

The first step of processing was to group SSL glosses according to which Swedish sentence they overlap most with. Second, glosses with certain marks were removed:

- Interrupted signs (marked `@&`).
- Gestures (marked `@g`).
- Incomprehensible signs (transcribed as `xxx`).

Finally, some marks were simply stripped from glosses, since they were not considered important to the current task.

- Signs performed with the non-dominant hand (marked `@nh`).
- Signs held with the non-dominant hand during production of signs with the dominant hand. The gloss of the held sign (following a `<>` symbol) was removed.
- Signs where the annotator was uncertain about which sign was used (marked `@xxx`).
- Signs where the annotator was uncertain about the correct gloss (marked `@zzz`).

In all, this is nearly identical to the procedure used by Sjons (2013, p. 14). Example 4.2 illustrates the output of the processing, with English glosses and translation added.

(4.2) *STÄMMA* *OCKSÅ* *PRO-1* *PERF* *BARN* *BRUKA* *SE*
be.correct also 1 PRF children usually watch
SAGA^TRÄD *PRO>närv*
Sagoträdet 2
'jag har ju barn också—brukar du se på Sagoträdet?'
'I also have children—do you watch Sagoträdet?'

The Swedish translations were tokenized and PoS-tagged with Stagger (Östling 2013), trained on the Stockholm-Umeå Corpus (SUC) (Ejerhed et al. 1992; Källgren 2006) and Stockholm Internet Corpus (SIC).[4]

[4] `http://www.ling.su.se/sic`

Table 4.8.: PoS tags in the SSLC, and their counterparts in SUC.

PoS	**SSLC**	**SUC**
Pronoun	PN	DT, HD, HP, HS, PS, PN
Noun	NN	NN, PM, UO
Verb	VB	PC, VB
Adverb	AB	AB, HA, IE, IN, PL
Numeral	RG	RG, RO
Adjective	JJ	JJ
Preposition	PP	PP
Conjunction	KN	KN, SN

4.3.2. Evaluation data

At the outset of the project, Carl Börstell and Lars Wallin manually assigned PoS tags to the 371 most frequent sign glosses in the corpus. This was used for initial annotation transfer experiments, and when the methods reached a certain level of maturity the remaining gloss types were automatically annotated, and the resulting list of 3,466 glosses manually corrected by Börstell and Wallin. Thus the initial goal of using annotation transfer to facilitate the PoS annotation was achieved, since all of the currently transcribed SSLC data now has manual annotations.

In order to evaluate the performance of the annotation transfer algorithms, I used this final set of 3,466 annotated types as a gold standard.

4.3.3. Tag set conversion

As previously mentioned, I used the eight PoS categories suggested by Ahlgren & Bergman (2006) for SSL. The Swedish side was tagged using the SUC tagset, whose core consists of 22 tags (Källgren 2006, p. 20). For direct tag projection and the tag translation priors in the circular generation model, the SUC tags were translated as in Table 4.8.

4.3.4. Task-specific tag constraints

Some sign glosses in the SSLC contain information that is relevant for their PoS.

- Proper nouns are marked with `@en`, and always receive the NOUN tag.
- Polysynthetic signs (Wallin 1994) are marked with `@p`, and always receive the VERB tag.
- Pronouns are glossed using `PRO-` or `POSS-`, and always receive the PRON tag.

Table 4.9.: Token-level PoS tagging accuracy, using direct projection from the final alignment (projection) or for the joint models, the sampled PoS tag variables (model). Note that in the former case, the PoS tag variables are ignored except during the alignment process. Figures given are averages ± standard deviation estimated over 64 randomly initialized experiments for each configuration.

	Types		**Tokens**	
	Project	**Model**	**Project**	**Model**
baseline	$58.4 \pm 0.5\%$	$12.2 \pm 0.6\%$	$75.3 \pm 0.7\%$	$10.8 \pm 3.6\%$
constraints	$58.4 \pm 0.4\%$	$60.7 \pm 0.4\%$	$75.1 \pm 0.8\%$	$58.1 \pm 1.0\%$
circular	$64.7 \pm 0.5\%$	$68.3 \pm 0.4\%$	$77.4 \pm 0.8\%$	$77.6 \pm 0.7\%$
circular + bigrams	$64.8 \pm 0.3\%$	$68.4 \pm 0.3\%$	$77.3 \pm 0.7\%$	$77.6 \pm 0.7\%$
circular + constraints	$69.1 \pm 0.4\%$	$77.1 \pm 0.3\%$	$79.7 \pm 0.6\%$	$78.7 \pm 0.6\%$
alternating	$56.8 \pm 0.5\%$	$56.8 \pm 0.5\%$	$73.1 \pm 0.9\%$	$73.4 \pm 0.8\%$
alternating + constraints	$64.1 \pm 0.6\%$	$75.7 \pm 0.4\%$	$76.2 \pm 0.9\%$	$77.1 \pm 0.7\%$

- Glosses whose names correspond to a Swedish lemma with an unambiguous PoS in the SALDO morphological lexicon (Borin & Forsberg 2009) always receive that tag.

All of these constraints except the last have very high precision but low recall. The last constraint assumes that the SSL signs whose glosses have names borrowed from Swedish words behave like these, which is not always the case. For instance, Swedish has a large open class of adjectives, whereas in SSL adjectives form a smaller, closed class. Apart from this, the constraint is rather accurate since signs have been disambiguated during the glossing process, so that instances of the same sign might have received different glosses depending on which PoS it was used as in a given instance.

4.3.5. Experimental results and analysis

In the experiments, different variations of the circular generation method were compared to the alternating alignment-annotation method. The particular implementation used in these experiments predates the one described in Section 4.1.3, deviating from it in two significant ways: no affix trees were used (this would be redundant, given the task-specific constraints), and only one alignment direction was used.

Table 4.9 shows the token-level accuracy for different models. Note that signs are assumed to belong to a single PoS, so that tags are assigned to types and token-level figures are derived from this by multiplication with the type frequency. To my surprise, the best result was obtained by using the circular model, with the alternating model close behind. It is also clear that using the task-specific tag constraints made a big contribution to these scores, and without these constraints the gap between the circular and

alternating models increased considerably. This outcome is interesting in light of evaluations on other data by myself for the New Testament (NT) corpus, and by Yarowsky & Ngai (2001) for the larger Canadian Hansards corpus (English-French), both of which speak in favor of the robust transfer method of Yarowsky & Ngai (2001) that is used in the alternating alignment-annotation algorithm.

One possible explanation for this result is the relatively small difference in accuarcy between the direct projections (first column of Table 4.9) and the models' PoS tags (second column). Whereas Yarowsky & Ngai (2001) reported an error rate of 24% for direct projection and 4% for robust transfer, there was virtually no difference when using the alternating alignment-annotation method on the SSLC. This is probably due to the data set itself, which was small (around 25,000 tokens) and consisted of unedited spontaneous conversation with plenty of disfluency.

Yarowsky & Ngai (2001) assume that it is possible to train a good bigram HMM tagger from projected data, which constitutes the only coupling between the word alignments and the final PoS tags. In contrast, the circular model samples each alignment link and PoS tag together, favoring consistency given a direct projection assumption.

Looking at the *circular+bigram* row of Table 4.9, we can see that introducing tag bigram dependencies into the circular model hurt token-level accuracy. The reason for this seems to be that the bigram model introduced a strong bias towards common PoS tags (nouns and verbs), which decreased token-level accuracy by mis-tagging some common pronouns and function words, but actually increased type-level accuracy (not shown) somewhat since these open-class tags were more common overall.

Perhaps the most important conclusion about annotation transfer that could be made from this experiment is that the choice of method depends crucially on the data at hand, and in some instances the simple circular generation model can be the best choice.

4.4. Lemmatization transfer

During the study described in Section 4.5, the need for a multilingual lemmatization or stemming tool emerged. As discussed in Section 2.3.6, such a tool can also be used to improve alignment performance by reducing data sparsity. Due to time constraints no proper evaluation has been undertaken, and the purpose of this section is mainly to provide essential background to Section 4.5. Nevertheless, an informal evaluation has shown that the method presented performs reasonably well for a variety of languages. Since simple concatenative morphology is assumed, however, phenomena like stem alterations or fusional morphology are not handled. Generalization to other types of morphology as well as a proper evaluation is left to future work.

The model that was used was a Bayesian segmentation model in the spirit of Goldwater (2007, Chapter 4), but with multilingual supervision akin to Yarowsky et al. (2001). In short, each target-language word form was "mirrored"[5] through the lemmas of the source language(s) it was aligned to, returning a set of word forms in the target language that were also translated by the same source language lemma(s). One of the problems

[5]Like the "semantic mirroring" of Dyvik (2005).

with most monolingual unsupervised morphology induction approaches is that, a priori, any two vocabulary items are potentially inflections of the same lexeme. This makes it plausible that *spe* is the stem of both *spend* and *speak*, even though the words are clearly unrelated. By using the mirror images from a lemmatized translation, it is possible to filter out nearly all of the similar but unrelated words, so that the algorithm can make stronger assumptions about the remaining candidates. For instance, the mirror image of *spends* might be: {*spend*, *spending*, *spends*, *consume*, *used*}. Given this set, it is straightforward to identify the stem *spend*, which in turn gives the suffixes *-ing* and *-s*.

For each target language word form w_i there is a latent variable s_i which represents the stem of w_i. These stems generate the mirror images of their respective word forms so that the probability of a word/stem pair i is:

$$P(w_i, s_i) = \prod_{w' \in M(w_i)} p(w'|s_i)$$

where

$$p(word|stem) = \begin{cases} (1-p_0)p(prefix)p(stem)p(suffix) & \text{if } stem \text{ is a substring of } word \\ p_0 & \text{otherwise} \end{cases}$$

and the three distributions over prefixes, suffixes and stems are categorical distributions with Dirichlet process priors, in these experiments with parameter $\alpha = 0.001$ for the first two distributions and $\alpha = 0.1$ for the stem distribution. The intuition between these parameter values is that stems form an "open" class of morphemes, while prefixes and suffixes are "closed" and contain relatively few members. Note that the segmentation is uniquely determined by the stem, by the simple rule that w' is split around the leftmost occurrence of s_i. If s_i does not occur in w', there is also a p_0 probability of generating an unrelated word form, in order to handle for instance synonymity and alignment errors. A value of $p_0 = 10^{-5}$ has turned out to give reasonable results. Note that the same word form can occur in several different mirror images, and thus has to be generated multiple times by the model. Consistency between multiple generations is encouraged (but not enforced) through the categorical distributions. Inference in the model is performed through ten iterations of Gibbs sampling, similar to Goldwater (2007, p. 43).

4.5. Lexical typology through multi-source concept transfer

While linguistic typology has traditionally focused on how structural or phonetic properties vary across languages, the field of lexical typology investigates the mapping between words and semantic domains across languages. For instance, the concepts of TREE and FIRE are expressed using different words in most of the world's languages, while a number of languages spoken on the Australian continent use the same word (see Figure 4.5). This is an instance of colexification in the sense of François (2008, p. 170):

> A given language is said to colexify two functionally distinct senses if, and only if, it can associate them with the same lexical form.

This definition intentionally circumvents the often difficult distinction between polysemy and semantic vagueness, and provides an operationalizable way to explore the lexical structure of languages without having to consider the actual forms of words.

Two different examples can be found in Figures 4.2 and 4.3, showing languages with STONE-MOUNTAIN and DIE-BLOOD colexification. Both cases are localized to one or a few regions, while almost unattested in the rest of the world. Each shape/shade combination represents a particular language family, according to the top-level classification of Hammarström et al. (2014). In the case of DIE-BLOOD, although a fairly large geographical area is covered, all of the languages are Sino-Tibetan and the actual word forms are similar.[6] This indicates either a genetic explanation or possibly borrowing. For STONE-MOUNTAIN, the situation is different. There are a few different areas (in central Africa, southern Africa, Australia, parts of South America) where this colexification is frequent, but in all these cases there is a broad representation of language families and word forms, which suggests that this is an areal phenomenon (or rather, several independent such phenomena).

Most of the colexifications discussed here have been previously studied. TREE-FIRE(-FIREWOOD) was discussed in depth for a large sample of Papuan and Australian languages by Hendery et al. (forthcoming). Aikhenvald (2009) similarly discussed EAT-DRINK-SMOKE colexification in Manambu (a Sepik language spoken in Papua New Guinea) and in other languages of the area. Urban (2012) covered patterns of colexifications for a large number of concepts and languages. Brown (2013b,a) has studied ARM-HAND and HAND-FINGER colexification and has made these data sets public. Other studies focus on how different parts of the color space are colexified (Kay & Maffi 2013a,b).

The traditional tools of lexical typology are lexica, word lists and, when available, human informants. Except for languages with well-structured digital lexica (which List et al. (2014) have utilized), this is a time-consuming manual process. Parallel texts have also been applied to this problem in the past; Wälchli & Cysouw (2012), for instance, use manually extracted examples from a subset of the New Testament in 101 different translations to explore the semantics of motion events.

My goal in this study is to automate the process of finding expressions of semantic concepts across a wide range of languages, by automatic word alignment from a number of source texts to each of the texts in the languages under study. To do this, semantic concepts are defined using word forms in the source languages. For instance, to find instances of the concept HAND given English, German and Chinese as source languages, one can (approximately) conclude that words linked to English *hand*, German *Hand* and Chinese 手 represent this concept. This information can then be used in exploring patterns of colexification. The method presented rests on a number of assumptions, which should be made explicit because they are sometimes violated:

- **Compositionality**: Idiomatic expressions like the English *to lend somebody a hand*, which does not involve the semantic concept HAND, are a potential source of error. To some extent this can be countered using multiple source languages, since

[6] There is only a single case attested outside this area: Sumerian, an extinct isolate.

only instances where the concept HAND is intended are likely to be consistently expressed with translation equivalents of the English *hand* across several languages.

- **Literal translation**: Since concepts that may be expressed with the same word naturally tend to be rather close semantically, it is often possible to convey roughly the same meaning using either of two concepts. For instance, where some translations have "he who has an ear" others might have "he who can hear," thereby making a comparison between the concepts EAR and TO HEAR difficult.

- **Word-to-word translation**: Since the underlying alignment algorithm is word-based, it becomes difficult to identify colexification when the concepts are not expressed with a single word. Exceptions to this assumption can be found in languages with noun incorporation, where the concept is expressed by a single morpheme inside a complex word, or, at the other extreme, in languages where the concept is expressed using a multi-word expression.

In order to explore the feasibility of this method for large-sample lexical typology, I evaluated the results against two existing databases: the WALS chapter on HAND-ARM (Brown 2013b) and the ASJP database for TREE-FIRE and STONE-MOUNTAIN (Wichmann et al. 2013).

4.5.1. Method

When defining concepts, it is desirable to avoid idiosyncrasies in particular languages or translations, in order find instances of each concept in the text that are likely to be translated consistently across languages. This suggests that it would be best to use a large and diverse sample of languages to define the concepts, but this ambition is somewhat hampered by practical concerns. In my experiments, two translations each in English, Swedish, Mandarin Chinese and Vietnamese were used as source texts. These languages are chosen because good supervised lemmatizers were available (English and Swedish) or because no such tool was needed due to the isolating nature of the languages (Chinese and Vietnamese). High-quality PoS tags were used for English and Swedish, which were used in the concept definitions to avoid some homographs, such as the English verb/noun pair *(to) stone*. Each concept was defined by specifying the lemma in a subset of the available languages, as well as (optionally) the PoS. In the case of common homographs that can not be readily separated from the intended lexeme, the form in a particular language was sometimes omitted entirely. This was mostly a problem in one of the Mandarin translations, where missing tone marks result in an excessive number of homographs for most monosyllabic words. Table 4.10 shows the definitions used for some concepts, where the Mandarin *shù* 'tree' is omitted because it has common homophones in the corpus.

The source texts were automatically word-aligned using the alternating alignment-annotation method described in Section 4.1.3. Any word that was aligned to at least k of the concept word forms per text (on average) was considered an instance of the concept, assuming that it made up a proportion of at least r of the total links from

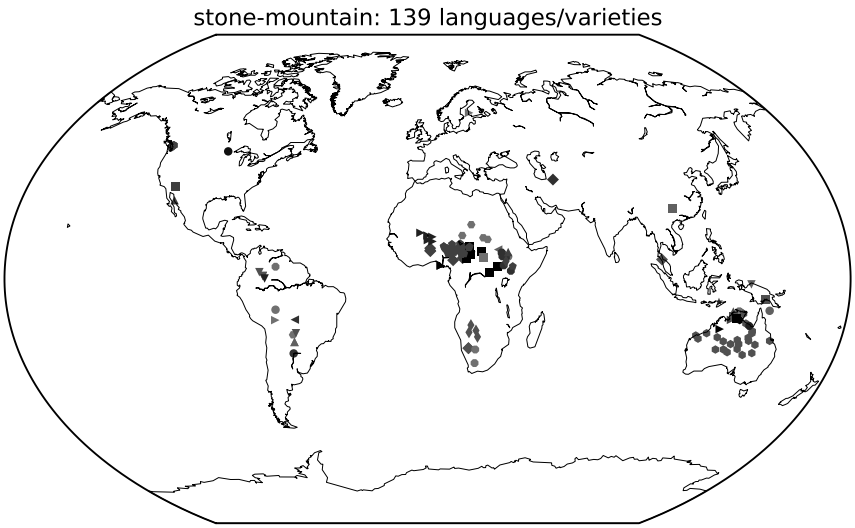


Figure 4.2.: Languages with STONE-MOUNTAIN colexification. Each combination of shape and shade represents a particular language family, according to the Glottolog classification (Hammarström et al. 2014). The purpose of this map is to illustrate the geographic and genealogical distribution of colexification, and the data was taken from ASJP (Wichmann et al. 2013) rather than from the method presented here.

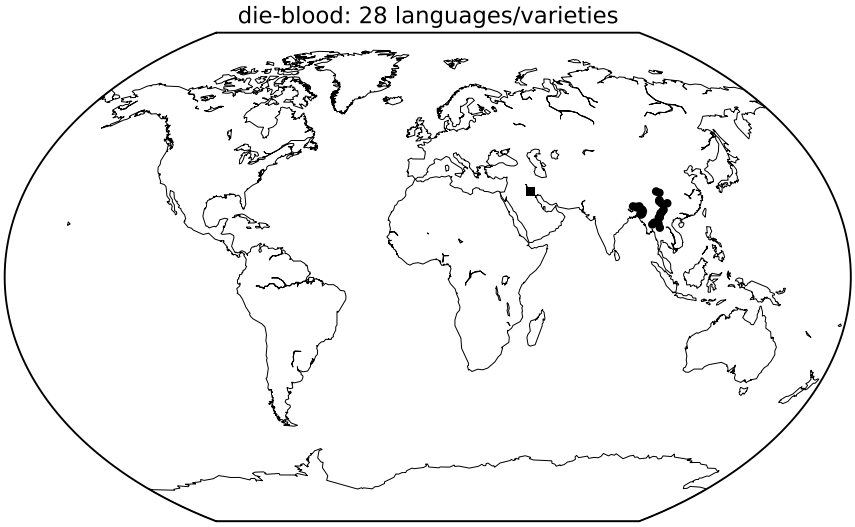


Figure 4.3.: Languages with DIE-BLOOD colexification. Each combination of shape and shade represents a particular language family, according to the Glottolog classification (Hammarström et al. 2014). The purpose of this map is to illustrate the geographic and genealogical distribution of colexification, and the data was taken from ASJP (Wichmann et al. 2013) rather than from the method presented here.

Table 4.10.: Example definitions of four concepts.

Language	MOUNTAIN	STONE	TREE	FIRE
English	mountain	stone	tree	fire
Swedish	berg	sten	träd	eld
Mandarin	shān	shítou	(shù)	huǒ

the concept word forms. The particular values used for k and r control the tradeoff between precision and recall, where high values of both favor precision, and low values favor recall. In other words, high values result in a very conservative classifier that only makes a guess when the number of instances found is large enough. Conversely, low values of k and r result in a classifier that makes guesses based on shaky evidence and is prone to misclassification. In the evaluation reported here, $k = 2$ and $r = 1/8$. Given more data a higher value of k would have been desirable in order to increase precision, but since there are so few instances of some of these concepts, this would decrease recall to unacceptable levels.

4.5.2. Evaluation

Although there was a large number of potential test cases, it turned out that most of them had to be discarded for one of the following reasons:

1. The concepts occur too rarely or not at all in the New Testament.
2. The concepts are too frequently collocated, like EAT and DRINK, so that they are difficult to separate during alignment.
3. The colexification occurs rarely or not at all in the language sample available.

There were three suitable colexification patterns which occurred in at least ten languages in the intersection between the evaluation set (WALS or ASJP) and the New Testament corpus: HAND-ARM, TREE-FIRE and STONE-MOUNTAIN.

The correctness of WALS and ASJP were assumed in my evaluation, but one should keep in mind that this is only approximately true. For instance, Japanese is classified as having identical words for HAND and ARM by Brown (2013b), but I have been informed by a linguist who is also a native speaker of Japanese that this appears to be a mistake in WALS (Yoko Yamazaki, p.c.). The algorithm correctly identified that 手 is most commonly used for HAND while 腕 is used for ARM. The reason seems to be that 手 is listed as an (obscure) translation of ARM in some dictionaries. If this is a general tendency in the WALS data, it would help to explain why HAND-ARM recall is so much lower than the other cases in Table 4.11.

These results are clearly not perfect, but are sufficient for identifying a likely set of candidates which can be explored in greater depth through other means. It can also

Table 4.11.: Agreement between algorithm and ASJP/WALS. L is the set of languages identified as colexifying a given concept by the algorithm, and G is the gold standard set. The gold standard consists of languages in *both* the New Testament corpus and the external data source (ASJP or WALS), the size of which is given in the Sample column.

Concepts	$\lvert L\rvert$	$\lvert L \cap G\rvert$	$\lvert G\rvert$	Sample	Precision	Recall
STONE-MOUNTAIN	24	9	12	821	38%	75%
TREE-FIRE	15	11	14	821	73%	79%
HAND-ARM	51	27	92	225	53%	29%

be sufficient for drawing preliminary conclusions, such as that the languages identified as colexifying TREE and FIRE are mostly spoken in Papua New Guinea and belong to different families, indicating an areal phenomenon (cf. Figures 4.4 and 4.5). The patterns discovered can then be explored using more precise—but also much more time-consuming—methods.

4.5.3. Limitations

The method that was used for this study has several apparent limitations. First, it is highly dependent on the particular text used as source material. If a concept is not present in the text, it is impossible to investigate it using this method. Even concepts that are present can be difficult to investigate, if they occur so rarely that it is difficult to identify them accurately, or if they frequently co-occur with other concepts. For instance, it is very difficult to study EAT-DRINK colexification with the New Testament since these are often expressed in the same sentences, making it difficult to differentiate between words expressing them, which in turn leads to a high rate of false positives. This particular problem could in some cases be alleviated by excluding verses where both concepts occur, but then there might not be a sufficient number of instances left to make a reliable match.

Furthermore, this method rests on the assumption that we can define a concept using lemmas from a few languages, and that we can identify equivalent words in other languages through parallel texts. There are many cases in which this assumption is broken. For instance, where one English version might have "eat," another has "take food," and this complicates automated studies of EAT-FOOD colexification.

Biases in the translations will also affect the result. With the New Testament, this results in very poor coverage for, among others, the native languages of Australia and North America. In the case of TREE-FIRE (Figure 4.4), this means that the method presented does not discover that the pattern actually extends beyond Papua, into the Australian continent (Figure 4.5, see also Hendery et al. (forthcoming)).

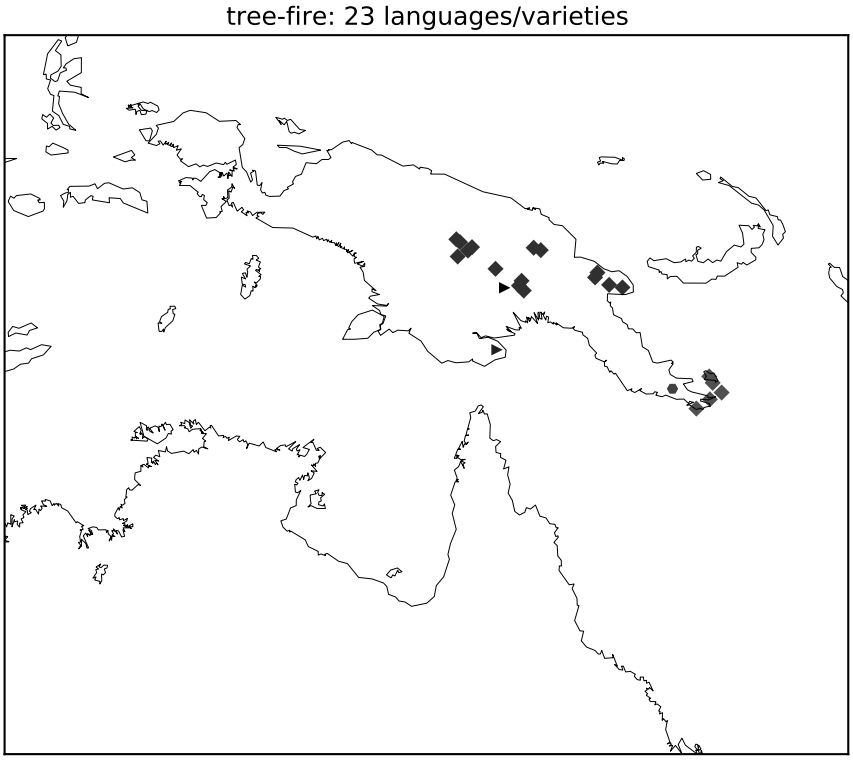


Figure 4.4.: Languages with TREE-FIRE colexification, according to the algorithm presented. Each combination of shape and shade represents a particular language family, according to the Glottolog classification (Hammarström et al. 2014). All languages are contained within the area of the map.

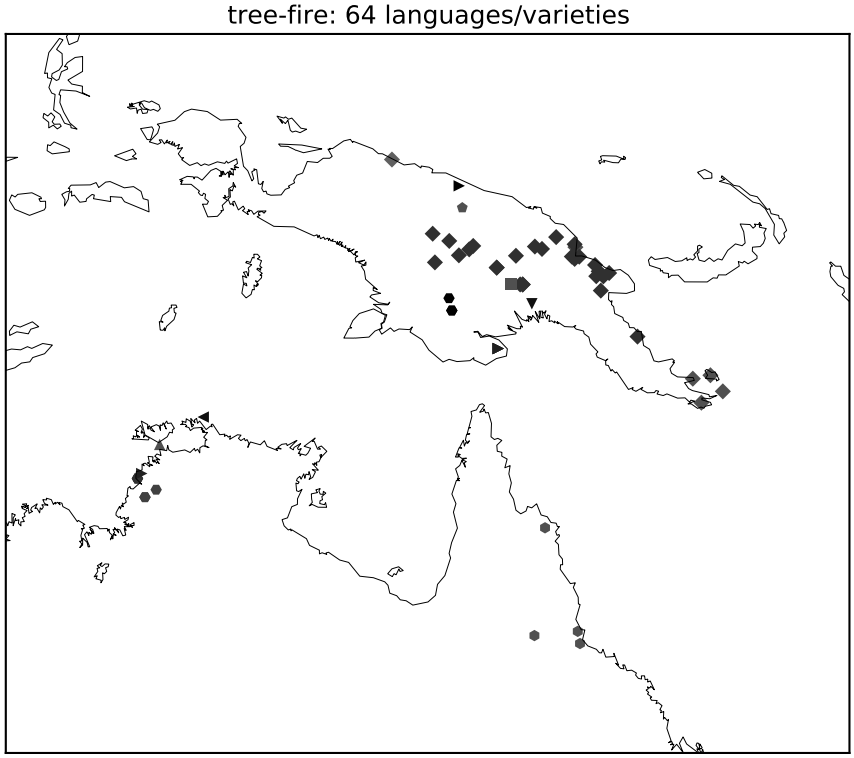


Figure 4.5.: Languages with TREE-FIRE colexification, according to ASJP (Wichmann et al. 2013). Each combination of shape and shade represents a particular language family, according to the Glottolog classification (Hammarström et al. 2014). Only a handful of languages are outside the area of the map, scattered around the world.

5. Multilingual word alignment

5.1. Interlingua alignment

Word alignment is normally defined between two languages, that is, *bitext alignment*. This is natural in many common applications such as single-source MT, but the recent appearance of parallel corpora with many languages and applications that go beyond MT leaves us with the question of how to align parallel texts of more than two languages.

Section 2.3.2 discusses some previous approaches that use information from more than two languages to perform word alignment. I introduce another approach, referred to here as *interlingua alignment* (Östling 2014).

The basic idea behind this family of methods is to learn a single abstract representation of the parallel text, to which all languages are aligned separately. A special case of interlingua alignment is the use of a bridge language, to which the texts in all other languages are aligned. This approach assumes that the bridge language text contains all relevant information in all the different translations, an assumption which is clearly too strong (see Figure 5.1).

Instead of using a fixed text as bridge language, the bridge language is *learned* along with alignments to all the translations in a parallel text. Ideally, this interlingua representation will then represent the information contained in *all* of the translations in the parallel text, which makes aligning to it easy (Figure 5.2).

I earlier presented a method for learning this Interlingua through Gibbs sampling of interlingua tokens (Östling 2014). The model essentially consists of n Dirichlet-multinomial IBM model 1 alignment pairs (as described in Section 2.5.3), one for each of the L languages being aligned. The interlingua is treated as the source language,

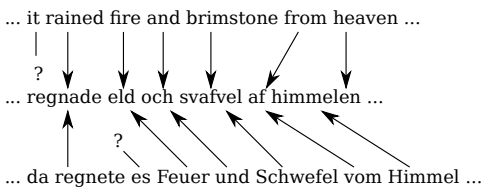


Figure 5.1.: Bridge language alignment. Note that the (Swedish) bridge language text does not include the dummy subject present in the other languages (English *it*, German *es*), which makes a satisfactory alignment impossible.

... it rained fire and brimstone from heaven ...
... regnade eld och svafvel af himmelen ...
... da regnete es Feuer und Schwefel vom Himmel ...

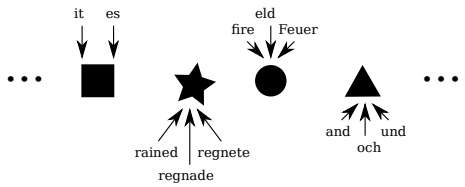


Figure 5.2.: Interlingua alignment (partial). Unlike the bridge language approach in Figure 5.1, the dummy subject is represented (as a square) so that the languages that use it can align to it.

assumed to generate all of the other L languages.

In addition to sampling the alignment variables a_j (Algorithm 2), the interlingua tokens (or *concepts*) e_i are also sampled, based on the current alignments. The effect of this is to make the interlingua tokens tend towards consistency with the words they are aligned to.

$$p(e_i = e|\boldsymbol{a}, \boldsymbol{e}_{-i}, \boldsymbol{f}) \propto p(e|\boldsymbol{e}_{-i}) \prod_l p(\boldsymbol{a}^{(l)}, \boldsymbol{f}^{(l)}|\boldsymbol{a}^{(l)}_{-i}, \boldsymbol{e}_{-i}, \boldsymbol{f}^{(l)}_{-i}) \tag{5.1}$$

where $p(e|\boldsymbol{e}_{-i})$ is a CRP prior[1] for concepts with concentration parameter α:

$$p(e|\boldsymbol{e}_{-i}) = \frac{1}{n+\alpha} \cdot \begin{cases} \alpha & \text{if } n_e = 0 \\ n_e & \text{if } n_e > 0 \end{cases} \tag{5.2}$$

The second factor describes the probability of adding, for each language l, a concept e with zero or more alignment links. For every interlingua concept e and language l, we have a Dirichlet-multinomial lexical translation distribution $p_t^{(l)}(f|e)$. Any given interlingua token e_i is aligned to a set of language tokens, which we need to consider when sampling e_i:

$$p(\boldsymbol{a}^{(l)}, \boldsymbol{f}^{(l)}|\boldsymbol{a}^{(l)}_{-i}, \boldsymbol{e}_{-i}, \boldsymbol{f}^{(l)}_{-i}) \propto \prod_{f:c(f)>0} \frac{\prod_{k=1}^{c(f)} \alpha_i + n_f + k - 1}{\prod_{k=1}^{\sum_f c(f)} (\sum_f \alpha_f + n_f) + k - 1} \tag{5.3}$$

[1] In hindsight, the general Pitman-Yor CRP would have been more suitable than the special case (Dirichlet process, with zero discount) I used previously (Östling 2014), since its parameters can be chosen so that the distribution of token frequencies more closely resembles the Zipf distribution seen in natural languages—and presumably the interlingua. See Section 2.4.5 for further discussion and references. In practice, this does not appear to matter much, since the second term of Equation (5.1) far outweighs the prior.

where $c(f) = |\{j : a_j = i \wedge f_j = f\}|$ is the number of times a particular target language type f is aligned to e_i. Since it is uncommon in practice with multiple tokens of the same type in a single sentence linked to one and the same interlingua token, $c(f)$ rarely exceeds 1.

The structure of this model is in fact similar to that of Snyder et al. (2009), except that the application of the latter is completely different in that it generates multilingual PoS tags, given a word-aligned parallel corpus. In that model, the multilingual PoS tags fill the same function as the interlingua tokens in my model, but instead of generating words directly, they assume multilingual tags generating (through fixed word alignment links) monolingual PoS tags, which in turn generate the words.

Computing Equation (5.1) requires $\mathrm{O}\big(L \cdot |E| \cdot |\boldsymbol{f}|\big)$ operations, where L is the number of languages, $|E|$ is the number of interlingua concepts, and $|\boldsymbol{f}|$ is the average number of tokens in the different target language texts $\boldsymbol{f}^{(l)}$. For the New Testament corpus described in Section 2.1, with 1,142 translations, this means roughly 10^{12} high-level operations for a single iteration of sampling $\boldsymbol{e}$.

Instead of assuming the interlingua $\boldsymbol{e}$ generates each translation $\boldsymbol{f}^{(l)}$, it is possible to reverse the alignment direction and make the opposite assumption: the interlingua is generated (independently) by the translations. This requires us instead to evaluate a large number of probabilities of the form $p_t^{(l)}(e|f)$, which can be sped up considerably by using the fact that $p_t^{(l)}(e|f) = \epsilon$, a constant, for most values of e (those that are never linked to f). Unfortunately, this variant of the model assigns high probability to solutions where all interlingua tokens are identical, which is clearly not a desirable solution.

5.1.1. Evaluation: Strong's numbers

For bitext alignment there are established evaluation metrics (Och & Ney 2003; Mihalcea & Pedersen 2003). These are based on the availability of gold-standard word linkage matrices for sentence pairs. With the interlingua alignment models described in this chapter, such an evaluation would be difficult. There are two conceivable ways of applying standard evaluation metrics to an interlingua alignment: evaluating pairwise alignments between individual languages, or evaluating pairwise alignments between the interlingua and each language.

5.1.1.1. Language-language pairwise alignments

The first way is to use (or generate, from whatever annotation is available) pairwise linkage matrices for each combination of languages, as well as pairwise alignments using the interlingua alignments as a bridge. There are, however, drawbacks to this approach: it is quadratic in the number of languages, and it is not obvious how to interpret and summarize the metrics obtained from all these pairwise alignments.

5.1.1.2. Interlingua-language pairwise alignments

Since the interlingua is constantly changing, it is unreasonable to demand a human-annotated gold standard for an interlingua-language bitext. However, the potential for a number of evaluations linear in the number of languages is appealing. I have previously described such a method for evaluating interlingua alignments of the New Testament corpus annotated with Strong's numbers (Östling 2014).

5.1.1.3. Clustering-based evaluation with Strong's numbers

Strong's numbers refer to the numbering in the King James Bible concordance of James Strong (1822–1894) and are used to map English words to the corresponding roots in the original texts. These numbers have later been added to other translations, and currently nine translations with Strong's numbers are available in the NT corpus. Cysouw et al. (2007) were the first to use these annotations to evaluate word alignments, although only for pairwise alignments.

We can abstractly view interlingua word alignment as a clustering problem. Each interlingua concept is a cluster, and all word tokens from a given language aligned to it are members of this cluster. Thus, we have one clustering for each language. Strong's numbers also define a clustering: each number is a cluster, and all tokens annotated with it are members of the cluster. Given this clustering view of the problem, we can now for each language compare the interlingua clustering with the clustering of Strong's numbers using standard clustering evaluation measures.

The Normalized Mutual Information (NMI) measure (Strehl & Ghosh 2003), also reinvented[2] as the V-measure by Rosenberg & Hirschberg (2007), is defined as:

$$\mathrm{NMI}(C, D) = \frac{2 \cdot I(C, D)}{H(C) + H(D)} \tag{5.4}$$

where $I(C, D)$ is the mutual information between clusterings C and D, and $H(C)$ is the entropy of clustering C. For a clustering C which contains clusters c_i, we can view C as a distribution, intuitively representing the probability that a randomly chosen element happens to be in cluster c_i:

$$p(c_i) = \frac{|c_i|}{n}$$

where n is the total numbers of elements. This allows the standard definition of entropy to be used:

$$H(C) = -\sum_i p(c_i) \log p(c_i)$$

[2] The equivalence was pointed out and proved by Becker (2011, pp 165–166). There are two versions of the measure, one of which uses the arithmetic mean (Equation (5.4)) and is equivalent to the V-measure, and another which uses the geometric mean $\sqrt{H(C)H(D)}$. The figures reported here and earlier (Östling 2014) use the arithmetic mean.

The joint probability of two clusters c_i and d_j from clusterings C and D over the same set of elements can be defined as:

$$p(c_i, d_j) = \frac{|c_i \cap d_j|}{n}$$

so that the mutual information can be defined as usual:

$$I(C, D) = -\sum_{i,j} p(c_i, d_j) \log \frac{p(c_i, d_j)}{p(c_i)p(d_j)}$$

Finally, there are some practical considerations when converting a text with Strong's numbers into a clustering. As discussed by Cysouw et al. (2007), not all words have Strong's number annotations, some have several, and the scopes of annotations are not given. They handle this by defining different types of agreement between alignments and Strong's numbers. While this offers a more precise picture, it comes at the expense of increased complexity compared to single-dimensional metrics. I bypass this problem by simply disregarding all tokens that do not have exactly one Strong's number. There are a number of theoretical objections one could have to this approach, maybe most importantly that it creates a bias towards evaluating "easy" words with one-to-one correspondences (between Greek and the language in question, at least). It is also not guaranteed to achieve full recall, since a number could potentially also refer to some preceeding word(s). In the end, the purpose of the evaluation is to compare the performance of different algorithms and parameters on the same data, which means that absolute figures matter less than relative ones, and a systematic bias can be tolerated.

Evaluation of word alignment is a difficult and controversial topic, for several reasons. As with most linguistic classification tasks, there is no obvious and universally applicable definition of what makes a "good" word alignment. Since word alignment is primarily an enabling technology, where the results are not used directly but rather as input for other tasks like SMT, the usefulness of evaluating word alignments on their own is questionable. Several authors have shown that common word alignment evaluation metrics correlate poorly with SMT evaluations (Vilar et al. 2006; Fraser & Marcu 2007; Holmqvist & Ahrenberg 2011). Nevertheless, unless one has a particular application in mind, there is little choice when exploring word alignment algorithms other than comparing to human-annotated alignments using some kind of similarity measure.

5.1.1.4. Bitext evaluation with Strong's numbers

In order to use the New Testament corpus for evaluating bitext alignments, it is useful to define a transformation from Strong's numbers to the common format with an alignment matrix with three-valued entries: no link, probable link, or sure link. There are two important complications in this conversion:

1. If multiple tokens with the same Strong's number occur in a verse, the alignment becomes ambiguous.

2. In the corpus annotations, the scope of a Strong's number is sometimes ambiguous. If one Greek lemma is translated using a multi-word expression, only the last word is annotated.

Therefore, a sequence of words without Strong's number annotations followed by an annotated word is ambiguous, as the annotation could apply to a (possibly empty) subset of the preceding words. To capture this uncertainty, *probable* links are added between all words in the two languages that might share the same number in this way. The first case, when a Strong's number occurs multiple times in a verse, is also handled by using probable links. Sure links are reserved for the case when a given Strong's number is only possible for one word per verse.

In practice, using this method results in about 2–3 probable links per token on average, but only around 0.2 sure links per token. This overgeneration of probable links and undergeneration of sure links impacts different evaluation measures in different ways. The sure-only and probable-only F-scores (F_S and F_P) will be unusually low under these conditions, while the AER and normal F-score are not severely affected.

5.2. Experiments

The first question to ask about the interlingua alignment model presented in this chapter is whether learning an interlingua results in better-quality word alignments, compared to simply picking a language and using that as an interlingua. In order to answer this question, I performed an experiment where the nine languages of the NT corpus containing Strong's number annotations were aligned. The interlingua representation was initialized in two separate experiments to either the English King James Version, or to a Mandarin Chinese translation (which was not among the nine translations to be aligned). Since the interlingua was constantly changing, the alignment variables (which are of primary interest) were not compatible between different samples, so I used simulated annealing (Section 2.5.4) rather than Rao-Blackwellization (Section 2.5.5).

To initialize the alignments, 200 alignment sampling iterations were performed with $\lambda = 1/\tau$ increasing linearly from 0 to 2, followed by two iterations with $\tau \to 0$ to find a locally optimal alignment to the initial interlingua (English or Chinese). Finally, 1000 iterations with $\tau = 1$ followed by two iterations with $\tau \to 0$ were performed with joint interlingua and alignment sampling. The learning curve of this last phase is shown in Figures 5.3 and 5.4, plotting the NMI of each language against the number of sampling iterations. Note that they start with a sharp drop (not shown in the figures, but see the left columns of Table 5.1), when sampling noise is introduced to the locally optimal alignments from the previous step. At the end there is a corresponding sharp increase, as we move to another (usually better) local optimum. The start and end points of these figures are given in Table 5.1, where we can see that in all but two cases there are improvements. The exceptions are for English texts, when English is used as the initial interlingua. Naturally, it is difficult to find a better representation for an English text, than the same (or very similar) English text. By comparing the English-initialized and Chinese-initialized interlingua, it is also clear that the latter has more room for

Table 5.1.: Normalized mutual information with respect to Strong's numbers, using alignment only (A) or joint alignment + interlingua learning (A+J), for models initialized using English or Mandarin.

	English		Mandarin	
	A	A+J	A	A+J
deu	0.817	**0.824**	0.708	**0.788**
eng	**0.854**	0.851	0.714	**0.800**
eng_2	**0.834**	0.833	0.708	**0.790**
fra	0.807	**0.816**	0.712	**0.783**
ind	0.774	**0.785**	0.710	**0.770**
ind_2	0.791	**0.803**	0.721	**0.786**
nld	0.839	**0.850**	0.724	**0.809**
por	0.807	**0.813**	0.709	**0.782**
rus	0.792	**0.800**	0.699	**0.772**

improvement. This is to be expected, since most of the evaluated languages are closely related to English. Given a sufficient number of iterations, the interlingua representations are guaranteed to be sampled from the same distribution. There are two reasons why this is not achieved in practice, resulting in the differences observed between the English-initialized and Chinese-initialized models: the insufficient number of sampling iterations (due to computational constraints), as well as the fact that the lengths of the interlingua sentences were chosen to be identical to those of the language used for initialization.

These runs each required about 350 hours to complete, on a 4-core Intel Core2 system running at 2 GHz. As can be seen in Figure 5.3, even 1,000 iterations was not enough for convergence. The computational intensity unfortunately limits the amount of experimentation that can be done, especially for large corpora such as the full NT corpus with its 1,142 translations.

5.3. Word order typology

Since the work of Greenberg (1963), word order features have played a central role in linguistic typology research. There is a great deal of variation across languages, and interesting interactions between different features that may hint at cognitive constraints in the processing of human language. A full theoretical discussion on word order typology is beyond the scope of this thesis, but the interested reader is referred to e.g. Dryer (2007) for an overview of the field.

I have applied the interlingua alignment method to investigate different word order features across the many languages of the New Testament corpus, by means of high-precision multi-source annotation transfer via the interlingua.

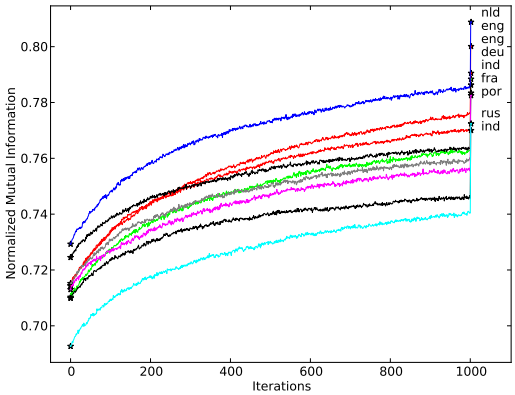


Figure 5.3.: Interlingua alignment model training, initialized with a Mandarin Chinese translation.

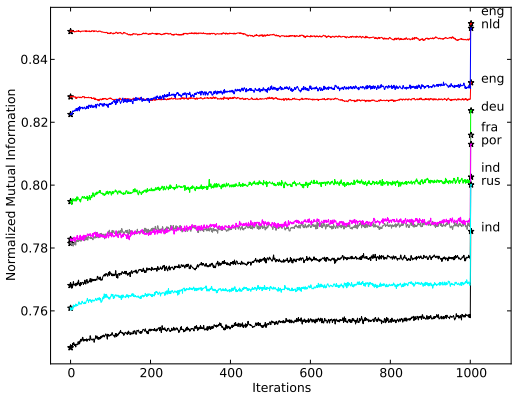


Figure 5.4.: Interlingua alignment model training, initialized with the English King James Version.

5.3.1. Method

The first step consisted of using supervised systems for annotating the source texts with Universal PoS Tags (Petrov et al. 2012) and dependency structure in the Universal Dependency Treebank format (McDonald et al. 2013). For PoS tagging, I used the Stanford Tagger (Toutanova et al. 2003) followed by a conversion step from the various language-specific tagsets to the "universal" tags of Petrov et al.. Next, I used the MaltParser dependency parser (Nivre et al. 2007) trained on the Universal Dependency Treebank using MaltOptimizer (Ballesteros & Nivre 2012).

From the source texts, PoS and dependency annotation was transferred to the interlingua representation. Since alignments were noisy and low recall was acceptable in this task, I used an aggressive filtering scheme: dependency links must have been transferred from at least 75% of source texts in order to be included. For PoS tags, which were only used to double-check grammatical relations and should not have impacted precision negatively, the majority tag among aligned words was used. Apart from compensating for noisy alignments and parsing errors, this method also helped to catch violations against the direct correspondence assumption (Hwa et al. 2002) by filtering out instances where different source texts use different constructions, favoring the most prototypical cases.

Finally, annotations were transferred directly from the interlingua to each of the texts in the corpus. This part of the process was the most prone to error, since alignment errors or violations of the direct correspondence assumption translate directly into incorrect annotations.

5.3.2. Data

The New Testament corpus was used, with ten English translations as source texts. Ideally more languages should be used, but at the time these experiments were performed I only had access to a preprocessing pipeline for English and German, and using some German translations in addition to the English ones did not lead to improved results.

5.3.3. Evaluation

In order to evaluate the results, data from World Atlas of Language Structures (WALS) (Dryer & Haspelmath 2013) was used. Only languages represented both in the New Testament corpus and the WALS data of a particular feature were used. A summary of this data is presented in Table 5.2. Although the New Testament corpus is a biased convenience sample, when comparing the values of the first column (for all languages in the WALS samples) with the second column (for languages in both the WALS samples and the New Testament corpus), we find that they are remarkably similar apart from a slight overrepresentation of verb-subject order in the New Testament subset.

Each word order feature was coded in terms of dependency relations, with additional constraints on the parts of speech that can be involved (e.g. subject-verb relations must be between a noun and a verb). The frequencies of all possible word orders for a feature were then counted, and for the purpose of evaluation the most common order was chosen as the algorithm's output. Although the relative frequencies of the different possible

Table 5.2.: Summary of the five features from WALS used in the experiment. The first column contains counts for all languages in the sample of each WALS feature, while the second column only includes the subset of languages which are *both* in WALS and the New Testament corpus.

81A: Order of Subject, Object and Verb (Dryer 2013e)		
523	142	SOV
465	132	SVO
88	40	VSO
23	8	VOS
10	3	OVS
4	1	OSV
167	32	No dominant order
1280	*358*	*Total*
82A: Order of Subject and Verb (Dryer 2013d)		
1115	283	SV
176	75	VS
97	25	No dominant order
1388	*383*	*Total*
83A: Order of Object and Verb (Dryer 2013c)		
655	168	OV
658	201	VO
92	16	No dominant order
1405	*385*	*Total*
85A: Order of Adposition and Noun Phrase (Dryer 2013b)		
541	155	Postpositions
479	154	Prepositions
8	0	Inpositions
55	18	No dominant order
28	3	No adpositions
1111	*330*	*Total*
87A: Order of Adjective and Noun (Dryer 2013a)		
341	97	Adjective-Noun
823	217	Noun-Adjective
101	34	No dominant order
4	0	Only internally-headed relative clauses
1269	*348*	*Total*

word orders were discarded for the sake of comparability with WALS, it was of course also possible to use these frequencies to estimate the level of freedom a given language allows in expressing a particular grammatical relation. Although an interesting topic, a detailed investigation into this is unfortunately not possible within the scope of this thesis.

In cases where there were multiple translations into a particular language, information was aggregated from all these translations into a single profile for the language. This was problematic in some cases, such as when a very large amount of time separated two translations and word order characteristics have evolved during that time. However, since the typical case was a single translation per language, and since WALS generally does not contain different data points for the different language varieties used, I leave the topic of historical change within a language to future research.

Not all languages and features can be easily classified into a particular word order. For instance, a language might lack adpositions altogether, or might not have any strong preference for either subject-verb or verb-subject order. These languages were excluded from the evaluation for the given feature. This meant that the task was made somewhat easier by excluding difficult cases, although in theory nothing prevents one from trying to detect and report these cases, too.

5.3.4. Results

Table 5.3 shows the agreement between the algorithm's output and the corresponding WALS chapter for each feature. The first thing to notice is the high level of agreement, even though the sample consisted mainly of languages unrelated to English, from which the dependency structure and PoS annotations were transferred. As expected, the lowest level of agreement is observed for WALS chapter 81A, which has a lower baseline since it allows six permutations of the verb, subject and object, whereas all the other features are binary. In addition, this feature requires that *two* dependency relations (subject-verb and object-verb) have been correctly transferred, which substantially reduces the number of relations available for comparison.

Since the accuracy of this method depended on the accuracy of the dependency link projections and word alignments, and the interlingua was fairly close to the English translation used to initialize it, one would expect the languages most different from English to be the most problematic. One way in which this can be observed is through looking at uncommon word orders. The one OSV language in the data (Nadëb, a Nadahup language from Brazil) was misidentified as SVO, probably due to the few projected transitive clauses (23). There were also three languages incorrectly classified as OSV, according to WALS, although one of them (Kaapor, a Tupian language from Brazil) does have OSV and SVO order according to the Ethnologue (Lewis et al. 2014). Similarly, the six languages identified as OVS were also done so incorrectly, according to WALS, although the Ethnologue indeed lists one of them (Barasana-Eduria, a Tucanoan language from Colombia) as an OVS language. It is unfortunate that the most exotic languages from the perspective of word order are also the ones most difficult to identify accurately.

Table 5.3.: Agreement between the algorithm and WALS. N is the number of languages that are both in the relevant WALS chapter and in the New Testament corpus, and where the WALS classification does indicate a particular ordering. For this reason, these counts are lower than the total counts in Table 5.2, which also include languages where a dominant ordering cannot be established. All features are binary except 81A, which can take six values.

Agreement	N	**Feature**
85.7%	342	81A: Order of Subject, Object and Verb
90.4%	376	82A: Order of Subject and Verb
96.4%	387	83A: Order of Object and Verb
95.1%	329	85A: Order of Adposition and Noun Phrase
88.0%	334	87A: Order of Adjective and Noun

The fact that sources sometimes differ as to the basic word order of a given language makes it evident that disagreement is not necessarily due to errors in the transfer process. Another example of this can be found when looking at the order of adjective and noun in some Romance languages (Spanish, Catalan, Portugese, French and Italian), which are all classified as having noun-adjective order (Dryer 2013a). It turned out that adjective-noun order in fact dominated in the NT with about two thirds of instances in all of these languages. This result was confirmed by manual inspection, which required a look at linguistic explanations for the discrepancy.[3] Both orders were very common, and arguments could be made for considering either (or none) of them as dominant. In favor of the noun-adjective order one could argue that it is normally less marked. Additionally, the common (but closed) class of quantifying adjectives uses adjective-noun order, so if these are excluded the case for noun-adjective is strengthened. On the other hand, adjective-noun was more common by a large margin, so if we take overall frequency in the New Testament as the deciding factor, it would be difficult to claim that noun-adjective is dominant. It is of course also important to note that the New Testament uses a rather formal register, and caution is warranted in generalizing from a single *doculect* (Cysouw & Good 2013).

5.3.5. Conclusions

The promising results from this study show that high-precision annotation transfer is a realistic way of exploring word order features in very large language samples, when a suitable parallel text is available. Although the WALS features on word order already use very large samples (over a thousand languages), using my method with the New Testament corpus contributes about 600 additional data points per feature—albeit with a small chance of error.

[3] Thanks to Francesca Di Garbo for helping with this.

The real strength of this method lies in the speed of execution, which allows the typologist to investigate previously unexplored features in seconds, using a sample of over a thousand languages. Moving beyond word order, there are numerous structural features of language that could be explored in a similar way, particularly if morpheme-based rather than word-based alignment is performed.

6. Conclusions

My main conclusion from the work presented in this thesis is that Bayesian models for word alignment offer an accurate, flexible and computationally efficient alternative to the EM-based algorithms that have been in use ever since the work of Brown et al. (1993). The simplicity of the Gibbs sampling algorithm used for inference allows the models to become more complex while keeping inference efficient and simple, and I have exploited this in two different ways: by improving bitext word alignment performance by performing Part of Speech (PoS) annotation transfer jointly with alignment, and by extending the bitext model to a multilingual word alignment model.

In addition to these developments related to word alignment as such, I have also explored several novel applications based on word-aligned parallel texts. First, I used joint alignment and annotation transfer in order to help providing the Swedish Sign Language Corpus with PoS tags, to my knowledge this is the first time automatic PoS annotation has been performed of a sign language corpus. Second, I have successfully applied my multilingual word alignment model to automatic investigations in word order typology, and in the field of lexical typology used the joint PoS transfer and word alignment algorithm for an automated colexification study. Apart from the new word alignment models presented, I would argue that my results further strengthen the case for the use of computational models in diverse areas of linguistic research.

In order to encourage adaptation of the methods I have developed and evaluated, the software implementation is available for download under a copyleft license.[1] For purposes of reproducibility, data and software from the experiments described in this thesis are archived at the Department of Linguistics, Stockholm University.

6.1. Future directions

Before speculating about the future of word alignment methods, we should take a step back and think about if there is one. Recent advances in neural network technology have led to models for Statistical Machine Translation (SMT) that directly map sentences from one language to another through a vector representation of the whole sentence, without the use of word alignments (Kalchbrenner & Blunsom 2013; Sutskever et al. 2014). As these new models are becoming competitive with traditional models based on word alignment, there is a real possibility that computing word alignments will become an obsolete problem, at least from the point of view of its historically most important application: SMT. In this thesis I presented several other applications based on word

[1] `http://www.ling.su.se/spacos`

alignments, but only the future can tell how many of those problems will continue to depend on word alignments.

That said, perhaps the most fundamental future project is how to move from *word alignment* to a more general alignment at several levels. The assumption that texts in different languages can be accurately linked at the word level is never entirely true, and the gap between this assumption and reality becomes even more acute when we try to align unrelated and structurally different languages. On the one hand, words are often too coarse to link accurately, because in most languages they tend to consist of multiple morphemes that could—or should—be linked individually rather than according to which word they are attached to. On the other hand, the word level may also be too fine-grained to be informative, when the correspondence is not between individual words but rather between whole constructions.

At a more technical level, we can ask which types of problems could and should be solved together with word alignment. I treated the case of PoS tagging in Chapter 4, and briefly touched upon lemmatization. There are however plenty of other candidates: syntactic parsing, morphological analysis, word sense disambiguation, and so on. A human translation of a text (let alone thousands of translations, as in the case of the New Testament) constitutes a fantastic amount of manual annotation, which we should not let go to waste.

A. Languages represented in the New Testament corpus

The following is a list of languages in the New Testament corpus described in Section 3.2.4. Languages are listed with 3-letter codes and names according to the ISO 639-3 standard, and are grouped according to the top-level classifications of Lewis et al. (2014).

Afro-Asiatic

Amharic (amh)
Bana (bcw)
Chadian Arabic (shu)
Dangaléat (daa)
Dawro (dwr)
Eastern Oromo (hae)
Gamo (gmv)
Gofa (gof)
Gude (gde)
Hausa (hau)
Hdi (xed)
Iraqw (irk)
Kafa (kbr)
Kambaata (ktb)
Kamwe (hig)
Kimré (kqp)
Konso (kxc)
Koorete (kqy)
Male (Ethiopia) (mdy)
Maltese (mlt)
Masana (mcn)
Matal (mfh)
Mbuko (mqb)
Merey (meq)
Mofu-Gudur (mif)
Muyang (muy)
Mwaghavul (sur)
Parkwa (pbi)
Sebat Bet Gurage (sgw)
Somali (som)
South Giziga (giz)
Tachelhit (shi)
Tamasheq (taq)
Tigrinya (tir)
Wandala (mfi)
Wolaytta (wal)
Zulgo-Gemzek (gnd)

Algic

Algonquin (alq)
Moose Cree (crm)
Severn Ojibwa (ojs)

Altaic

Azerbaijani (aze)
Gagauz (gag, 2 translations)
Halh Mongolian (khk)
Kalmyk (xal)
Kara-Kalpak (kaa)
Karachay-Balkar (krc)
Kazakh (kaz)
Khakas (kjh)
Kirghiz (kir)
Kumyk (kum)
Russia Buriat (bxr)
Southern Altai (alt)
Tatar (tat)
Tuvinian (tyv)
Uighur (uig, 2 translations)
Uzbek (uzb)

Arai (Left May)

Ama (Papua New Guinea) (amm)

Arauan

Paumarí (pad)

Australian

Burarra (bvr)
Djambarrpuyngu (djr)
Kuku-Yalanji (gvn)
Wik-Mungkan (wim)

Austro-Asiatic

Eastern Bru (bru)
Parauk (prk)
Vietnamese (vie, 4 translations)

Austronesian

'Auhelawa (kud)
Achinese (ace)

Agusan Manobo (msm)
Agutaynen (agn)
Alangan (alj)
Alune (alp)
Arifama-Miniafia (aai)
Arosi (aia)
Balantak (blz)
Balinese (ban)
Bambam (ptu)
Banggai (bgz)
Batad Ifugao (ifb)
Batak Angkola (akb)
Batak Dairi (btd)
Batak Karo (btx)
Batak Simalungun (bts)
Batak Toba (bbc)
Biatah Bidayuh (bth)
Bima (bhp)
Bola (bnp)
Bolinao (smk)
Botolan Sambal (sbl)
Brooke's Point Palawano (plw)
Bugawac (buk)
Buginese (bug)
Buhid (bku)
Bunama (bdd)
Bwanabwana (tte)
Caribbean Javanese (jvn)
Cebuano (ceb, 2 translations)
Central Bikol (bcl)
Central Dusun (dtp)
Central Malay (pse)
Central Sama (sml)
Chamorro (cha)
Cotabato Manobo (mta)
Da'a Kaili (kzf)
Dawawa (dww)
Dobu (dob)
Duri (mvp)
Eastern Tawbuid (bnj)
Fijian (fij)
Gapapaiwa (pwg)
Gilbertese (gil)
Gorontalo (gor)
Halia (hla)
Hanunoo (hnn)
Hawaiian (haw)
Hiligaynon (hil)
Hote (hot)
Iamalele (yml)
Iban (iba)
Iduna (viv)
Iloko (ilo)
Indonesian (ind, 9 translations)
Iraya (iry)
Itawit (itv)
Iwal (kbm)
Jarai (jra)
Javanese (jav)
Kahua (agw)
Kambera (xbr)
Kankanaey (kne)
Kapingamarangi (kpg)
Kara (Papua New Guinea) (leu)
Keapara (khz)
Keley-I Kallahan (ify)
Kinaray-A (krj)
Koronadal Blaan (bpr)
Kuanua (ksd)
Kwara'ae (kwf)
Lampung Api (ljp)
Ledo Kaili (lew)
Lote (uvl)
Ma'anyan (mhy)
Madak (mmx)
Madurese (mad)
Mag-antsi Ayta (sgb)
Mainstream Kenyah (xkl)
Makasar (mak)
Malagasy (mlg)
Malay (individual language) (zlm, 4 translations)
Mamasa (mqj)
Manam (mva)
Mangga Buang (mmo)
Manggarai (mqy)
Mangseng (mbh)
Maori (mri)
Mapos Buang (bzh)
Maranao (mrw)
Marik (dad)
Marshallese (mah)
Maskelynes (klv)
Matigsalug Manobo (mbt)
Mayoyao Ifugao (ifu)
Mbula (mna)
Mekeo (mek)
Mengen (mee)
Mentawai (mwv)
Minangkabau (min)
Misima-Panaeati (mpx)
Molima (mox)
Mongondow (mog)
Motu (meu)
Muna (mnb)
Mutu (tuc, 2 translations)
Muyuw (myw)
Nakanai (nak)
Napu (npy)
Nehan (nsn)
Ngaju (nij)
Nias (nia)
North Tanna (tnn)
Nyindrou (lid)
Obo Manobo (obo)
Owa (stn)
Paicî (pri)
Pamona (pmf)
Pampanga (pam)
Pangasinan (pag)
Paranan (prf)
Patep (ptp)
Patpatar (gfk)
Plateau Malagasy (plt)
Ramoaaina (rai)
Sabu (hvn)
Samoan (smo)
Sangir (sxn)

Saposa (sps)
Sarangani Blaan (bps)
Sasak (sas)
Seimat (ssg)
Sinaugoro (snc)
Sio (xsi)
Sissano (sso, 2 translations)
Southwest Tanna (nwi, 2 translations)
Suau (swp)
Sundanese (sun)
Sursurunga (sgz)
Tagalog (tgl, 2 translations)
Takia (tbc)
Tangoa (tgp)
Tawala (tbo)
Termanu (twu)
Timugon Murut (tih)
Tinputz (tpz)
Toraja-Sa'dan (sda)
Tungag (lcm)
Tuwali Ifugao (ifk)
Uab Meto (aoz)
Ubir (ubr)
Uma (ppk)
Uripiv-Wala-Rano-Atchin (upv)
Waima (rro)
Waray (Philippines) (war)
Western Bukidnon Manobo (mbb)
Western Penan (pne)
Wuvulu-Aua (wuv)
Yabem (jae)

Aymaran

Central Aymara (ayr, 2 translations)

Barbacoan

Awa-Cuaiquer (kwi)
Chachi (cbi)
Colorado (cof)

Border

Amanab (amn)
Waris (wrs)

Cahuapanan

Chayahuita (cbt)

Cariban

Akawaio (ake)
Apalaí (apy)
Bakairí (bkq)
Galibi Carib (car)
Hixkaryána (hix)
Macushi (mbc)
Patamona (pbc)
Wayana (way)

Chibchan

Border Kuna (kvn)
Bribri (bzd)
Buglere (sab)
Cabécar (cjp)
Central Tunebo (tuf)
Ngäbere (gym)
San Blas Kuna (cuk)
Teribe (tfr)

Chipaya-Uru

Chipaya (cap)

Chocoan

Epena (sja)
Northern Emberá (emp)
Woun Meu (noa)

Constructed language

Esperanto (epo)

Creole

Baba Malay (mbf)
Belize Kriol English (bzj)
Bislama (bis)
Eastern Maroon Creole (djk)
Haitian (hat)
Hawai'i Creole English (hwc)
Jamaican Creole English (jam)
Krio (kri)
Kriol (rop)
Morisyen (mfe)
Nigerian Pidgin (pcm)
Pijin (pis)
Saint Lucian Creole French (acf)
Sango (sag)
Saramaccan (srm)
Sea Island Creole English (gul)
Sranan Tongo (srn)
Tok Pisin (tpi)

Dravidian

Kannada (kan)
Malayalam (mal)

East Bird's Head-Sentani

Manikion (mnx)
Meyah (mej)
Moskona (mtj)

East Geelvink Bay

Bauzi (bvz)

East New Britain

Qaqet (byx)

Eastern Trans-Fly

Bine (bon)
Wipi (gdr)

Eskimo-Aleut

Eastern Canadian Inuktitut (ike)
Kalaallisut (kal)
North Alaskan Inupiatun (esi)
Northwest Alaska Inupiatun (esk)

Eyak-Athabaskan

Carrier (crx)
Dogrib (dgr)
Gwich'in (gwi)
Southern Carrier (caf)
Western Apache (apw)

Guajiboan

Cuiba (cui)
Guahibo (guh)
Guayabero (guo)

Guaykuruan

Kadiwéu (kbc)
Mocoví (moc)
Pilagá (plg)
Toba (tob)

Harákmbut

Amarakaeri (amr)

Hmong-Mien

Hmong Daw (mww)

Huavean

San Mateo Del Mar Huave (huv)

Indo-European

Afrikaans (afr, 5 translations)
Ancient Greek (to 1453) (grc, 2 translations)
Armenian (hye)
Awadhi (awa)
Breton (bre)
Bulgarian (bul)
Caribbean Hindustani (hns)
Catalan (cat, 2 translations)
Church Slavic (chu)
Croatian (hrv, 2 translations)
Czech (ces, 2 translations)
Danish (dan, 3 translations)
Dari (prs)
Dutch (nld, 5 translations)
English (eng, 10 translations)
Faroese (fao)
Fiji Hindi (hif)
French (fra, 6 translations)
German (deu, 12 translations)
Gujarati (guj)
Hindi (hin)
Icelandic (isl)
Iranian Persian (pes)
Irish (gle)
Italian (ita, 2 translations)
Latin (lat)
Latvian (lav, 2 translations)
Lithuanian (lit)
Macedonian (mkd)
Maithili (mai)
Marathi (mar)
Middle English (1100-1500) (enm, 2 translations)
Modern Greek (1453-) (ell, 2 translations)
Northern Kurdish (kmr)
Norwegian Bokmål (nob)
Norwegian Nynorsk (nno, 2 translations)
Ossetian (oss)
Plautdietsch (pdt)
Polish (pol, 3 translations)
Portuguese (por, 8 translations)
Romanian (ron)
Russian (rus, 3 translations)
Serbian (srp)
Sindhi (snd)
Sinte Romani (rmo)
Slovak (slk)
Slovenian (slv)
Spanish (spa, 11 translations)
Swabian (swg)
Swedish (swe, 6 translations)
Tajik (tgk)
Tosk Albanian (als, 2 translations)
Ukrainian (ukr)
Vlax Romani (rmy, 2 translations)
Welsh (cym, 2 translations)

Iroquoian

Cherokee (chr)

Japonic

Japanese (jpn)

Jean

Apinayé (apn)
Kaingang (kgp)
Kayapó (txu)
Xavánte (xav)

Jicaquean

Tol (jic)

Jivaroan

Achuar-Shiwiar (acu)
Aguaruna (agr)
Huambisa (hub)
Shuar (jiv)

Karajá

Karajá (kpj)

Khoisan

Nama (Namibia) (naq)

Language isolate

Basque (eus, 3 translations)
Camsá (kbh)
Candoshi-Shapra (cbu)
Chiquitano (cax)
Cofán (con)
Korean (kor)
Kuot (kto)
Rikbaktsa (rkb)
Sulka (sua)
Ticuna (tca)
Urarina (ura)
Waorani (auc)
Yuracare (yuz)

Maipurean

Ajyíninka Apurucayali (cpc)
Apurinã (apu)
Asháninka (cni)
Ashéninka Pajonal (cjo)
Caquinte (cot)
Garifuna (cab)
Ignaciano (ign, 2 translations)
Machiguenga (mcb)
Nomatsiguenga (not)
Parecís (pab)
Piapoco (pio)
Pichis Ashéninka (cpu)
Tereno (ter)
Trinitario (trn)
Ucayali-Yurúa Ashéninka (cpb)
Wapishana (wap)
Wayuu (guc)
Yanesha' (ame)
Yine (pib)
Yucuna (ycn)

Mapudungu

Mapudungun (arn)

Matacoan

Iyojwa'ja Chorote (crt)
Maca (mca)
Wichí Lhamtés Güisnay (mzh)
Wichí Lhamtés Nocten (mtp)

Maxakalian

Maxakalí (mbl)

Mayan

Achi (acr)
Aguacateco (agu)
Chol (ctu, 2 translations)
Chortí (caa)
Chuj (cac, 2 translations)
Huastec (hus, 2 translations)
Ixil (ixl, 2 translations)
K'iche' (quc, 2 translations)
Kaqchikel (cak, 7 translations)
Kekchí (kek)
Lacandon (lac)
Mam (mam, 4 translations)
Mopán Maya (mop)
Popti' (jac, 2 translations)
Poqomchi' (poh, 2 translations)
Q'anjob'al (kjb)
Tabasco Chontal (chf)
Tektiteko (ttc)
Tojolabal (toj)
Tz'utujil (tzj, 2 translations)
Tzotzil (tzo, 4 translations)
Uspanteco (usp)
Western Kanjobal (knj)
Yucateco (yua)

Misumalpan

Mayangna (yan)
Mískito (miq, 2 translations)

Mixe-Zoquean

Coatlán Mixe (mco)
Francisco León Zoque (zos)
Highland Popoluca (poi)
Isthmus Mixe (mir)
Juquila Mixe (mxq)
Mazatlán Mixe (mzl)
Tlahuitoltepec Mixe (mxp)
Totontepec Mixe (mto)

Mosetenan

Tsimané (cas)

Nambiquaran

Southern Nambikuára (nab)

Niger-Congo

Adamawa Fulfulde (fub)
Adele (ade)
Adioukrou (adj)
Akoose (bss)
Anufo (cko)
Bafia (ksf)
Bafut (bfd)
Bambara (bam)
Bekwarra (bkv)
Bembe (bmb)
Bete-Bendi (btt)
Bimoba (bim)
Bissa (bib)
Boko (Benin) (bqc)
Bokobaru (bus)
Bomu (bmq)
Buamu (box)
Buli (Ghana) (bwu)
Bulu (Cameroon) (bum)
Busa (bqp)
Cameroon Mambila (mcu)
Cerma (cme)
Chopi (cce)
Chumburung (ncu)
Dan (dnj)
Deg (mzw)
Delo (ntr)
Denya (anv)
Digo (dig)
Dii (dur)
Ditammari (tbz)
Djimini Senoufo (dyi)
Doyayo (dow)
Duruma (dug)
Dyula (dyu)
Eastern Karaboro (xrb)
Ekajuk (eka)
Ewe (ewe)
Farefare (gur, 2 translations)
Fon (fon)
Gen (gej)
Gikyode (acd)
Giryama (nyf)
Gitonga (toh)
Gogo (gog)
Gokana (gkn)
Gourmanchéma (gux)
Guinea Kpelle (gkp)
Gusii (guz)
Hanga (hag)
Haya (hay)
Hehe (heh)
Igbo (ibo)
Irigwe (iri)
Ivbie North-Okpela-Arhe (atg)
Izere (izr)
Jola-Fonyi (dyo)
Jola-Kasa (csk)
Jukun Takum (jbu)
Kabiyè (kbp)
Kagulu (kki)
Kako (kkj)
Kasem (xsm)
Kenyang (ken)
Kikuyu (kik)
Kim (kia)
Kinyarwanda (kin)
Konkomba (xon)
Konni (kma)
Kono (Sierra Leone) (kno)
Koongo (kng)
Koonzime (ozm)
Kouya (kyf)
Kuanyama (kua)
Kukele (kez)
Kuo (xuo)
Kuranko (knk)
Kusaal (kus)
Kutep (kub)
Kuwaa (blh)
Kuwaataay (cwt)
Laari (ldi)
Lama (Togo) (las)
Lelemi (lef)
Lenje (leh)
Lobi (lob)
Lozi (loz)
Lukpa (dop)
Lyélé (lee)
Machame (jmc)
Mada (Nigeria) (mda)
Malawi Sena (swk)
Malba Birifor (bfo)
Mamara Senoufo (myk)
Mampruli (maw)
Mandinka (mnk)
Masaaba (myx)
Mbunda (mck)
Mende (Sierra Leone) (men)
Miyobe (soy)
Moba (mfq)
Mochi (old)
Moro (mor)
Mossi (mos, 2 translations)
Mumuye (mzm)
Mundani (mnf)
Mwani (wmw)
Mündü (muh)
Nafaanra (nfr)
Nande (nnb)
Nawdm (nmz)
Ndogo (ndz)
Ndonga (ndo)
Ngangam (gng)
Ngiemboon (nnh)
Nigeria Mambila (mzk)
Nigerian Fulfulde (fuv)
Nilamba (nim)
Ninzo (nin)
Nkonya (nko)
Nomaande (lem)
Noone (nhu)
Northern Dagara (dgi)

Northern Grebo (gbo)
Northwest Gbaya (gya)
Ntcham (bud)
Nyabwa (nwb)
Nyakyusa-Ngonde (nyy)
Nyanja (nya, 2 translations)
Nyankole (nyn)
Nyaturu (rim)
Nyoro (nyo)
Obolo (ann)
Paasaal (sig)
Pedi (nso, 2 translations)
Plapo Krumen (ktj)
Pokomo (pkb)
Pular (fuf)
Rundi (run)
Saamia (lsm)
Sekpele (lip)
Selee (snw)
Shona (sna)
Sissala (sld)
South Fali (fal)
South Ndebele (nbl)
Southern Birifor (biv)
Southern Bobo Madaré (bwq)
Southern Kisi (kss)
Southern Nuni (nnw)
Southern Samo (sbd)
Southwest Gbaya (gso)
Suba (sxb)
Supyire Senoufo (spp)
Susu (sus)
Swahili (individual language) (swh)
Swati (ssw)
Tampulma (tpm)
Tharaka (thk)
Tikar (tik)
Timne (tem)
Toro So Dogon (dts)
Toura (Côte d'Ivoire) (neb)
Tsikimba (kdl)
Tsishingini (tsw)
Tsonga (tso)
Tswana (tsn)
Tumbuka (tum)
Tumulung Sisaala (sil)
Tupuri (tui)
Twi (twi)
Vagla (vag)
Venda (ven)
Vengo (bav)
Vunjo (vun)
Vute (vut)
Waama (wwa)
West-Central Limba (lia)
Wolof (wol)
Wè Northern (wob)
Xaasongaxango (kao)
Xhosa (xho)
Yalunka (yal)
Yamba (yam)
Yocoboué Dida (gud)
Yoruba (yor)
Zemba (dhm)
Zulu (zul)

Nilo-Saharan

Adhola (adh)
Alur (alz)
Avokaya (avu)
Bedjond (bjv)
Datooga (tcc)
Gulay (gvl)
Jur Modo (bex)
Karamojong (kdj)
Kenga (kyq)
Kumam (kdi)
Kupsabiny (kpz)
Lango (Uganda) (laj)
Luo (Kenya and Tanzania) (luo)
Luwo (lwo)
Ma'di (mhi)
Mabaan (mfz)
Markweeta (enb)
Mbay (myb)
Murle (mur)
Ndo (ndp)
Ngambay (sba)
Northeastern Dinka (dip)
Nuer (nus)
Sabaot (spy)
Sar (mwm)
Shilluk (shk)
Southwestern Dinka (dik)
Teso (teo)
Uduk (udu)
Zarma (dje)

North Bougainville

Rotokas (roo)

North Caucasian

Avaric (ava)
Chechen (che)
Tabassaran (tab)

Otomanguean

Amatlán Zapotec (zpo)
Atatláhuca Mixtec (mib)
Ayautla Mazatec (vmy)
Cajonos Zapotec (zad)
Central Mazahua (maz)
Chayuco Mixtec (mih)
Chichicapan Zapotec (zpv)
Chiquihuitlán Mazatec (maq)
Choapan Zapotec (zpc)
Coatecas Altas Zapotec (zca)
Coatzospan Mixtec (miz)
Comaltepec Chinantec (cco)
Copala Triqui (trc)
Diuxi-Tilantongo Mixtec (xtd)
Eastern Highland Chatino

(cly)
Eastern Highland Otomi (otm)
Estado de México Otomi (ots)
Guerrero Amuzgo (amu)
Huautla Mazatec (mau)
Isthmus Zapotec (zai)
Jalapa De Díaz Mazatec (maj)
Jamiltepec Mixtec (mxt)
Lachixío Zapotec (zpl)
Lalana Chinantec (cnl)
Lealao Chinantec (cle)
Magdalena Peñasco Mixtec (xtm)
Mezquital Otomi (ote)
Miahuatlán Zapotec (zam)
Mitla Zapotec (zaw)
Mixtepec Zapotec (zpm)
Nopala Chatino (cya)
Ocotepec Mixtec (mie)
Ocotlán Zapotec (zac)
Ozolotepec Zapotec (zao)
Ozumacín Chinantec (chz)
Palantla Chinantec (cpa)
Peñoles Mixtec (mil, 2 translations)
Pinotepa Nacional Mixtec (mio)
Querétaro Otomi (otq)
Quioquitani-Quierí Zapotec (ztq)
Quiotepec Chinantec (chq)
Rincón Zapotec (zar)
San Jerónimo Tecóatl Mazatec (maa, 2 translations)
San Juan Atzingo Popoloca (poe)
San Juan Colorado Mixtec (mjc)
San Juan Guelavía Zapotec (zab)
San Marcos Tlalcoyalco Popoloca (pls)
San Martín Itunyoso Triqui (trq)
San Miguel El Grande Mixtec (mig)
San Pedro Amuzgos Amuzgo (azg)
Santa María Quiegolani Zapotec (zpi)
Santo Domingo Albarradas Zapotec (zas)
Silacayoapan Mixtec (mks)
Sochiapam Chinantec (cso)
Southern Puebla Mixtec (mit)
Southern Rincon Zapotec (zsr)
Tabaa Zapotec (zat)
Tataltepec Chatino (cta)
Tenango Otomi (otn)
Tepetotutla Chinantec (cnt)
Tepeuxila Cuicatec (cux)
Teutila Cuicatec (cut)
Texmelucan Zapotec (zpz)
Tezoatlán Mixtec (mxb)
Usila Chinantec (cuc)
Western Highland Chatino (ctp)
Yalálag Zapotec (zpu)
Yatee Zapotec (zty)
Yatzachi Zapotec (zav)
Yosondúa Mixtec (mpm)
Zoogocho Zapotec (zpq)

Paezan

Guambiano (gum)
Páez (pbb)

Panoan

Capanahua (kaq)
Cashibo-Cacataibo (cbr)
Cashinahua (cbs)
Chácobo (cao)
Matsés (mcf)
Sharanahua (mcd)
Shipibo-Conibo (shp)
Yaminahua (yaa)

Pauwasi

Karkar-Yuri (yuj)

Pidgin

Hiri Motu (hmo)

Puinavean

Cacua (cbv)
Nadëb (mbj)

Quechuan

Ayacucho Quechua (quy)
Cajamarca Quechua (qvc)
Cañar Highland Quichua (qxr)
Chimborazo Highland Quichua (qug)
Cusco Quechua (quz)
Eastern Apurímac Quechua (qve)
Huallaga Huánuco Quechua (qub)
Huamalíes-Dos de Mayo Huánuco Quechua (qvh)
Huaylas Ancash Quechua (qwh)
Huaylla Wanca Quechua (qvw)
Imbabura Highland Quichua (qvi)
Inga (inb)
Lambayeque Quechua (quf)
Margos-Yarowilca-Lauricocha Quechua (qvm)

Napo Lowland Quechua (qvo)
North Bolivian Quechua (qul)
North Junín Quechua (qvn)
Northern Conchucos Ancash Quechua (qxn)
Northern Pastaza Quichua (qvz)
Panao Huánuco Quechua (qxh)
San Martín Quechua (qvs)
South Bolivian Quechua (quh)
Southern Conchucos Ancash Quechua (qxo)
Southern Pastaza Quechua (qup)
Tena Lowland Quichua (quw)

Ramu-Lower Sepik

Aruamu (msy)

Senagi

Angor (agg)

Sepik

Abau (aau)
Alamblak (amp)
Ambulas (abt, 2 translations)
Hanga Hundi (wos)
Iatmul (ian)
Kwanga (kwj)
Kwoma (kmo)
Mende (Papua New Guinea) (sim)
Saniyo-Hiyewe (sny)
Sepik Iwam (iws)
Yessan-Mayo (yss, 2 translations)

Sino-Tibetan

Achang (acn)
Akha (ahk)
Angami Naga (njm)
Bawm Chin (bgr)
Burmese (mya)
Falam Chin (cfm)
Haka Chin (cnh)
Hakka Chinese (hak)
Kachin (kac)
Lahu (lhu)
Lushai (lus)
Mandarin Chinese (cmn, 2 translations)
Maru (mhx)
Matu Chin (hlt)
Mün Chin (mwq)
Ngawn Chin (cnw)
Siyin Chin (csy)
Tedim Chin (ctd)
Thado Chin (tcz)
Zotung Chin (czt)
Zou (zom)

South Bougainville

Naasioi (nas)

South-Central Papuan

Tabo (knv, 2 translations)

Tacanan

Cavineña (cav)
Ese Ejja (ese)
Tacana (tna)

Tarascan

Purepecha (tsz)

Tequistlatecan

Highland Oaxaca Chontal (chd)

Torricelli

Au (avt, 2 translations)
Bukiyip (ape)
Bumbita Arapesh (aon)
Kamasau (kms)
Mufian (aoj, 2 translations)
Olo (ong)

Totonacan

Coyutla Totonac (toc)
Highland Totonac (tos)
Huehuetla Tepehua (tee)
Papantla Totonac (top)
Pisaflores Tepehua (tpp)
Tecpatlán Totonac (tcw)
Tlachichilco Tepehua (tpt)
Upper Necaxa Totonac (tku)
Xicotepec De Juárez Totonac (too)

Trans-New Guinea

Agarabi (agd)
Alekano (gah)
Amele (aey)
Aneme Wake (aby)
Angaataha (agm)
Angal Heneng (akh)
Angguruk Yali (yli)
Anjam (boj)
Ankave (aak)
Awa (Papua New Guinea) (awb)
Awiyaana (auy)
Barai (bbb)
Bargam (mlp)
Baruya (byr)

Benabena (bef)
Biangai (big)
Bimin (bhl)
Binumarien (bjr)
Borong (ksr)
Chuave (cjv)
Dadibi (mps)
Daga (dgz)
Dano (aso)
Dedua (ded)
East Kewa (kjs)
Ese (mcq)
Ewage-Notu (nou)
Faiwol (fai)
Fasu (faa)
Folopa (ppo)
Fore (for)
Girawa (bbr)
Golin (gvf)
Guhu-Samane (ghs)
Gwahatike (dah)
Huli (hui)
Imbongu (imo)
Inoke-Yate (ino)
Ipili (ipi)
Iyo (nca)
Kalam (kmh, 2 translations)
Kamula (xla)
Kanasi (soq)
Kanite (kmu)
Kein (bmh)
Keyagana (kyg)
Kobon (kpw, 2 translations)
Komba (kpf)
Korafe-Yegha (kpr)
Kosena (kze)
Kube (kgf)
Kuman (kue, 2 translations)
Kunimaipa (kup)
Kyaka (kyc)
Mape (mlh)
Mauwake (mhl)
Melpa (med)
Mian (mpt)
Mountain Koiali (kpx)
Nabak (naf)
Nalca (nlc)
Namiae (nvm)
Ngalum (szb)
Nii (nii)
Nobonob (gaw)
North Tairora (tbg)
Numanggang (nop)
Oksapmin (opm)
Orokaiva (okv)
Rawa (rwo, 2 translations)
Safeyoka (apz)
Salt-Yui (sll)
Samberigi (ssx)
Selepet (spl)
Siane (snp, 3 translations)
Siroi (ssd)
Somba-Siawari (bmu)
South Tairora (omw)
Suena (sue)
Telefol (tlf)
Timbe (tim)
Tuma-Irumu (iou)
Umanakaina (gdn)
Usan (wnu)
Usarufa (usa)
Waffa (waj)
Wantoat (wnc)
Weri (wer)
West Kewa (kew)
Wiru (wiu)
Yareba (yrb)
Yau (Morobe Province) (yuw)
Yaweyuha (yby)
Yongkom (yon)
Yopno (yut)
Zia (zia)
Ömie (aom)

Tucanoan

Barasana-Eduria (bsn)
Carapana (cbc)
Cubeo (cub)
Desano (des)
Guanano (gvc)
Koreguaje (coe)
Macuna (myy)
Piratapuyo (pir)
Secoya (sey)
Siona (snn)
Siriano (sri)
Tatuyo (tav)
Tucano (tuo)
Tuyuca (tue)
Waimaha (bao)

Tupian

Aché (guq)
Eastern Bolivian Guaraní (gui)
Guajajára (gub)
Guarayu (gyr)
Kaiwá (kgk)
Kayabí (kyz)
Mbyá Guaraní (gun)
Mundurukú (myu)
Paraguayan Guaraní (gug)
Sateré-Mawé (mav)
Sirionó (srq)
Tenharim (pah)
Urubú-Kaapor (urb)
Western Bolivian Guaraní (gnw)

Uralic

Eastern Mari (mhr)
Erzya (myv)
Estonian (est, 2 translations)
Finnish (fin, 3 translations)
Hungarian (hun)

Komi-Zyrian (kpv)
Northern Sami (sme)

Uto-Aztecan

Central Huasteca Nahuatl (nch)
Central Tarahumara (tar)
Eastern Huasteca Nahuatl (nhe)
El Nayar Cora (crn, 2 translations)
Guerrero Nahuatl (ngu)
Highland Puebla Nahuatl (azz)
Hopi (hop)
Huichol (hch)
Lowland Tarahumara (tac)
Mayo (mfy)
Michoacán Nahuatl (ncl)
Northern Oaxaca Nahuatl (nhy)
Northern Paiute (pao)
Northern Puebla Nahuatl (ncj)
Northern Tepehuan (ntp)
Southeastern Puebla Nahuatl (npl)
Southeastern Tepehuan (stp)
Tetelcingo Nahuatl (nhg)
Tohono O'odham (ood)
Western Huasteca Nahuatl (nhw)
Yaqui (yaq)
Zacatlán-Ahuacatlán-Tepetzintla Nahuatl (nhi)

West Papuan

Galela (gbi)
Tabaru (tby)
Tobelo (tlb)
Yawa (yva)

Witotoan

Bora (boa)
Minica Huitoto (hto)
Muinane (bmr)
Murui Huitoto (huu)

Yaguan

Yagua (yad)

Yanomaman

Sanumá (xsu)
Yanomámi (wca)

Yele-West New Britain

Pele-Ata (ata)
Yele (yle)

Zamucoan

Ayoreo (ayo)
Chamacoco (ceg)

Zaparoan

Arabela (arl)

Svensk sammanfattning

Introduktion

I den här avhandlingen närmar jag mig ett antal till synes väldigt olika problem: att hitta ordklasser i svenskt teckenspråk och 1 001 andra språk runt världen, att undersöka ordföljd i alla dessa språk, och att ta reda på om de gör skillnad på händer och armar. Det gemensamma temat som förenar dessa olika ämnen är metoden: *ordlänkning* i parallelltexter. Det här är en av de uppgifter som verkar busenkla för en nybörjare, och djävulskt svåra för oss som har försökt att programmera en dator att utföra dem. Problemet är detta: givet översatta meningar på olika språk, markera vilka ord som motsvarar varandra i dessa språk.

Bortsett från tillämpningarna ovan kommer mycket av den här avhandlingen behandla utveckling och utforskning av kärnmetoderna för ordlänkning, speciellt det nya fältet om Bayesianska modeller för ordlänkning med algoritmer av typen MCMC (eng. *Markov Chain Monte Carlo*) för inferens, i synnerhet Gibbs-algoritmen (eng. *Gibbs sampling*) (DeNero et al. 2008; Mermer & Saraçlar 2011; Gal & Blunsom 2013). Trots att tidigare studier har visat att MCMC-metoder är ett lockande alternativ till de EM-baserade (eng. Expectation-Maximization) metoderna som oftast används, är de få tidigare studierna relativt begränsade och jag såg ett behov av en bredare studie om Bayesianska ordlänkningsmodeller. I ett nötskal gör Bayesianska modeller det möjligt att enkelt vikta lösningen mot vad som är lingvistiskt troligt, till exempel genom att minska sannolikheten för ett överdrivet stort antal postulerade översättningar för varje ord, eller genom att öka sannolikheten för en realistisk frekvensdistribution för interlinguabegreppen som används i den flerspråkiga ordlänkningsalgoritm som presenteras i kapitel 5. Givet att man väljer lämpliga sannolikhetsfördelningar, kan Gibbs-algoritmen enkelt användas trots att modellens sannolikhetsfunktion är väldigt komplex.

Mina huvudsakliga forskningsfrågor kan sammanfattas som följer:

1. Vad kännetecknar MCMC-algoritmer för ordlänkning, hur ska de tillämpas i praktiken, och hur är de jämfört med andra metoder?

2. Hur kan man utföra ordlänkning i massivt parallella korpusar med hundratals eller tusentals språk?

3. Hur kan ordlänkade parallellkorpusar användas för att genomföra undersökningar inom lingvistisk typologi?

4. Kan ordlänkning och annotationsöverföring utföras samtidigt för att förbättra noggrannheten för båda uppgifterna?

Bidrag

Mina bidrag i den här avhandlingen är av olika typer: algoritmer, tillämpningar och utvärderingar. Dessa sammanfattas här, med referenser till vilka av de ovanstående forskningsfrågorna som besvaras.

Algoritmer

De främsta innovationerna i algoritmväg finns i kapitel 5, där en metod för flerspråkig ordlänkning genom ett interlingua presenteras (forskningsfråga 2), och i kapitel 4 där ordlänking och ordklassöverföring utförs gemensamt (forskningsfråga 4).

Flerspråklig ordlänkning är ett nytt område, där min metod utgör ett sätt att utnyttja den informationsrikedom som finns i massivt parallella texter, alltså texter som är översatta till hundratals eller tusentals språk. Medan normala ordlänkningsalgoritmer antar att texterna som ska länkas är fasta, försöker min algoritm i stället att samtidigt lära sig både en mellanrepresentation, ett *interlingua*, och länkar från denna till vart och ett av de många språken. Detta interlingua syftar till att utifrån de olika översättningarna så nära som möjligt approximera den semantiska information som finns i parallelltexten, och de ”ord” som används i representationen ska representera tvärspråkliga begrepp som överlag ligger nära betydelsen hos de ord som används i de olika språken.

Tidigare arbeten har visat dels att ordlänkning kan användas för att överföra lingvistisk annotation mellan språk där det finns parallelltexter, dels att sådan annotation kan användas för att ge en mer precis ordlänkning. Min nya algoritm som presenteras i kapitel 4 kan göra båda sakerna samtidigt, vilket leder till bättre ordlänkning när bara ett av språken som länkas har lingvistiska annotationer. Detta är en ganska typisk situation i praktiken, eftersom de allra flesta av världens språk saknar språkspecifika verktyg för lingvistisk annotation.

Tillämpningar

Ordlänkning i sig är inte speciellt spännande för de flesta. Med detta i åtanke har jag försökt att tillämpa mina ordlänkningsalgoritmer på några valda problem från olika områden inom lingvistik och språkteknologi. De flesta av tillämpningarna skulle också vara möjliga att genomföra med andra ordlänknigsmetoder än de som jag utforskar, och syftet är inte främst att testa mina egna algoritmer (för det, se nedan), utan att inspirera andra att använda ordlänkade parallelltexter i sin forskning.

Min algoritm för samtidig ordlänkning och annotationsöverföring (kapitel 4, forskningsfråga 4) har jag använt i avsnitt 4.3 för att föra över ordklassannotation till transkriberat svenskt teckenspråk i Institutionen för lingvistiks svenska teckenspråkskorpus. Detta är första gången teckenspråk har annoterats med ordklasser med hjälp av automatiska verktyg. Dessutom har annotationen möjliggjort ny grundforskning om svenskt teckenspråk.

Forskningsfråga 3 berörs främst i två tillämpningar. Till att börja med har jag använt den flerspråkiga länkningsalgoritmen i kapitel 5 för att undersöka ordföljd i de 1 001

språken där jag har tillgång till översättningar av Nya Testamentet, genom att projicera dependensstruktur först till interlinguarepresentationen och sedan vidare till de olika översättningarna (avsnitt 5.3). Genom att jämföra med WALS-databasen (Dryer & Haspelmath 2013) kan man se att svaren man får stämmer för mellan 86% och 96% av språken, beroende på vilket fenomen man tittar på.

Vidare har jag använt min algoritm för samtidig ordlänkning och annotationsöverföring för att undersöka vilka språk som *kolexifierar* olika begrepp, alltså uttrycker dem med samma ord (avsnitt 4.5). Till exempel kan metoden upptäcka att ELD och TRÄD uttrycks med samma ord i ett antal obesläktade språk på Papua med omnejd, ett fynd som bekräftas av andra studier och ASJP-databasen Wichmann et al. (2013).

Utvärderingar

Kapitel 3 och 4 innehåller noggranna utvärderingar för ett antal språkpar, och visar att Bayesianska ordlänkningsmodeller är konkurrenskraftiga för en bredare mängd språk än vad som tidigare visats (forskningsfråga 1). Dessutom innebär de en stark baslinje för de fortsatta experimenten som presenteras i dessa kapitel. Jag undersöker ett antal frågor som tidigare forskning lämnat obesvarade: vilken effekt hierarkiska Pitman-Yordistributioner har på ordlänkningsalgoritmen jämfört med den mer beräkningsmässigt effektiva icke-hirerarkiska Dirichlet-distributionen, hur väl en explicit Gibbs-sampler fungerar för ordlänkning, och hur olika sätt att initiera modellen påverkar resultatet. Vidare publicerar jag genomgående den statistik som krävs för varje experiment för att beräkna de många utvärderingsmått som används. Jag hoppas att det här kan inspirera andra till att göra samma sak, vilket skulle leda till att resultaten från ordlänkningsstudier blir lättare att jämföra i framtiden.

Slutsatser

Min huvudsakliga slutsats från arbetet som presenteras i avhandlingen är att Bayesianska modeller för ordlänkning erbjuder ett precist, flexibelt och beräkningsmässigt effektivt alternativ till de EM-baserade algoritmer som använts ända sedan Brown et al. (1993). Enkelheten i Gibbs-algoritmen tillåter modellerna att bli mer komplexa samtidigt som inferens förblir effektiv och enkel, och jag har utnyttjat detta på två olika sätt: genom att förbättra noggrannheten vid ordlänkning av bitexter genom att samtidigt föra över ordklassannotationer, samt genom att utöka bitextmodellen till en flerspråkig länkningsmodell.

För att uppmuntra andra att använda metoderna jag har utvecklat och utvärderat, finns min implementation tillgänglig för nedladdning under en copyleft-licens.[1] Dessutom finns data och mjukvara för experimenten som beskrivs arkiverade på Institutionen för lingvistik på Stockholms universitet.

[1] `http://www.ling.su.se/spacos`

Framtida arbete

Innan vi börjar spekulera om framtiden för ordlänkningsmetoder, bör vi ta ett steg tillbaka och fundera på om det faktiskt finns någon. Den senaste tidens framsteg inom neuronnätsteknik har lett till modeller för statistisk maskinöversättning (SMT) som direkt översätter meningar i ett språk till ett annat genom vektorrepresentationer för hela meningen, utan att använda ordlänkar (Kalchbrenner & Blunsom 2013; Sutskever et al. 2014). I takt med att de nya översättningsmodellerna börjar kunna konkurrera med traditionella ordlänkningsbaserade modeller, finns det en verklig möjlighet att beräkning av ordlänkar blir ett ointressant problem, i alla fall sett från dess mest inflytelserika tillämpning: maskinöversättning. I den här avhandlingen har jag presenterat flera andra tillämpningar baserade på ordlänkar, men bara framtiden kan utvisa hur många av dessa problem som kommer att fortsätta vara beroende av ordlänkar.

Med detta sagt, är det kanske viktigaste framtida projektet att gå vidare från *ordlänkning* till en mer generell länkning på flera nivåer. Antagandet att texter på olika språk kan länkas på ordnivå är aldrig helt sant, och klyftan mellan detta antagande och verkligheten växer ytterligare när vi försöker länka obesläktade och strukturellt olika språk. Å ena sidan är ord ofta för grova enheter för att kunna länkas på ett bra sätt, eftersom de brukar bestå av flera morfem som i de flesta språk kan, eller borde, länkas individuellt snarare än beorende på vilket ord de tillhör. Å andra sidan kan ordnivån också vara för detaljerad, när motsvarigheten gäller hela konstruktioner snarare än enskilda ord.

På en mer teknisk nivå kan vi fråga oss vilka typer av problem som kan och borde lösas tillsammans med ordlänkning. Jag har behandlat fallen med ordklasstaggning i kapitel 4 och i förbifarten nämnt lemmatisering, men det finns en uppsjö av andra kandidater: syntaktisk parsning, morfologisk analys, orddisambiguering, och så vidare. En översättning gjord av en människa, för att inte tala om de tusentals översättningar som finns av Nya Testamentet, innebär ett enormt annotationsarbete som vi inte bör låta gå till spillo.

Bibliography

Ahlgren, I. & Bergman, B. (2006). Det svenska teckenspråket. In *Teckenspråk och teckenspråkiga: kunskaps- och forskningsöversikt*, volume 2006:29 of *Statens offentliga utredningar (SoU)* (pp. 11–70). Ministry of Health and Social Affairs.

Aikhenvald, A. Y. (2009). The linguistics of eating and drinking. In J. Newman (Ed.), *'Eating', 'drinking' and 'smoking': A generic verb and its semantics in Manambu* (pp. 91–108). Amsterdam: John Benjamins.

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, 50(1-2), 5–43.

Aswani, N. & Gaizauskas, R. (2005). A hybrid approach to align sentences and words in English-Hindi parallel corpora. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ParaText '05 (pp. 57–64). Stroudsburg, PA, USA: Association for Computational Linguistics.

Ayan, N. F. & Dorr, B. J. (2006). A maximum entropy approach to combining word alignments. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference* (pp. 96–103). New York City, USA: Association for Computational Linguistics.

Ballesteros, M. & Nivre, J. (2012). MaltOptimizer: An optimization tool for MaltParser. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12 (pp. 58–62). Stroudsburg, PA, USA: Association for Computational Linguistics.

Beal, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, University of London.

Becker, H. (2011). *Identification and Characterization of Events in Social Media*. PhD thesis, Columbia University.

Besag, J. (2004). *Mathematical Foundations of Speech and Language Processing*, chapter An Introduction to Markov Chain Monte Carlo Methods, (pp. 247–270). Springer: New York City.

Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *The Annals of Mathematical Statistics*, 18(1), 105–110.

Blunsom, P., Cohn, T., Goldwater, S., & Johnson, M. (2009). A note on the implementation of hierarchical Dirichlet processes. In *Proceedings of the ACL-IJCNLP*

2009 Conference Short Papers, ACLShort '09 (pp. 337–340). Stroudsburg, PA, USA: Association for Computational Linguistics.

Bojar, O. & Prokopová, M. (2006). Czech-English word alignment. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1236–1239). Genova, Italy: ELRA.

Borin, L. (2000). You'll take the high road and I'll take the low road: Using a third language to improve bilingual word alignment. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00 (pp. 97–103). Stroudsburg, PA, USA: Association for Computational Linguistics.

Borin, L. & Forsberg, M. (2009). All in the family: A comparison of SALDO and WordNet. In *NODALIDA 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies* (pp. 7–12). Odense, Denmark.

Börstell, C., Mesch, J., & Wallin, L. (2014). Segmenting the Swedish Sign Language corpus: On the possibilities of using visual cues as a basis for syntactic segmentation. In O. Crasborn, E. Efthimiou, E. Fotinea, T. Hanke, J. Hochgesang, J. Kristoffersen, & J. Mesch (Eds.), *Beyond the Manual Channel. Proceedings of the 6th Workshop on the Representation and Processing of Sign Languages* (pp. 7–10). Reykjavík, Iceland: ELRA.

Brown, C. H. (2013a). Finger and hand. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Brown, C. H. (2013b). Hand and arm. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, 18(4), 467–479.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2), 263–311.

Cherry, C. & Lin, D. (2003). A probability model to improve word alignment. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (pp. 88–95). Sapporo, Japan: Association for Computational Linguistics.

Cherry, C. & Lin, D. (2006a). A comparison of syntactically motivated word alignment spaces. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* Trento, Italy.

Cherry, C. & Lin, D. (2006b). Soft syntactic constraints for word alignment through discriminative training. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions* (pp. 105–112). Sydney, Australia: Association for Computational Linguistics.

Chung, T. & Gildea, D. (2009). Unsupervised tokenization for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (pp. 718–726). Singapore: Association for Computational Linguistics.

Croft, W. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press.

Cysouw, M., Biemann, C., & Ongyerth, M. (2007). Using Strong's Numbers in the Bible to test an automatic alignment of parallel texts. *STUF - Language Typology and Universals*, 60(2), 158–171.

Cysouw, M. & Good, J. (2013). Languoid, doculect, and glossonym: Formalizing the notion 'language'. *Language Documentation & Conservation*, 7, 331–359.

Cysouw, M. & Wälchli, B. (2007). Parallel texts: Using translational equivalents in linguistic typology. *STUF - Language Typology and Universals*, 60(2), 95–99.

Dahl, Ö. (2007). From questionnaires to parallel corpora in typology. *STUF - Language Typology and Universals*, 60(2), 172–181.

Das, D. & Petrov, S. (2011). Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11 (pp. 600–609). Stroudsburg, PA, USA: Association for Computational Linguistics.

De Vries, L. (2007). Some remarks on the use of Bible translations as parallel texts in linguistic research. *STUF - Language Typology and Universals*, 60(2), 148–157.

Dejean, H., Gaussier, E., Goutte, C., & Yamada, K. (2003). Reducing parameter space for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond - Volume 3*, HLT-NAACL-PARALLEL '03 (pp. 23–26). Stroudsburg, PA, USA: Association for Computational Linguistics.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1), 1–38.

DeNero, J., Bouchard-Côté, A., & Klein, D. (2008). Sampling alignment structure under a Bayesian translation model. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 314–323). Honolulu, Hawaii: Association for Computational Linguistics.

Dryer, M. S. (2007). Word order. In T. Shopen (Ed.), *Language Typology and Syntactic Description*, volume I chapter 2, (pp. 61–131). Cambridge University Press.

Dryer, M. S. (2013a). Order of adjective and noun. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dryer, M. S. (2013b). Order of adposition and noun phrase. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dryer, M. S. (2013c). Order of object and verb. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dryer, M. S. (2013d). Order of subject and verb. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dryer, M. S. (2013e). Order of subject, object and verb. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Dryer, M. S. & Haspelmath, M. (2013). *The World Atlas of Language Structures Online*. `http://wals.info`.

Dyer, C., Chahuneau, V., & Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 644–648). Atlanta, Georgia: Association for Computational Linguistics.

Dyvik, H. (2005). Translations as a semantic knowledge source. In *Proceedings of the Second Baltic Conference on Human Language Technologies* (pp. 27–38). Tallinn: Institute of Cybernetics, Tallinn University of Technology & Institute of the Estonian Language.

Ejerhed, E., Källgren, G., Wennstedt, O., & Åström, M. (1992). *The Linguistic Annotation System of the Stockholm-Umeå Corpus Project*. Technical report, Department of Linguistics, University of Umeå.

Emmorey, K. (2003). *Perspectives on Classifier Constructions in Sign Langauges*. Mahwah, NJ: Lawrence Erlbaum and Associates.

Eyigöz, E., Gildea, D., & Oflazer, K. (2013). Simultaneous word-morpheme alignment for statistical machine translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 32–40). Atlanta, Georgia: Association for Computational Linguistics.

Filali, K. & Bilmes, J. (2005). Leveraging multiple languages to improve statistical MT word alignments. In *IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 92–97). San Juan: IEEE.

Fortescue, M. (1984). *West Greenlandic*. Croom Helm Descriptive Grammars. London: Croom Helm.

Fox, H. J. (2002). Phrasal cohesion and statistical machine translation. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02 (pp. 304–3111). Stroudsburg, PA, USA: Association for Computational Linguistics.

Franz, A., Kumar, S., & Brants, T. (2013). 1993-2007 United Nations parallel text. Linguistic Data Consortium, Philadelphia.

François, A. (2008). Semantic maps and the typology of colexification: Intertwining polysemous networks across languages. In M. Vanhove (Ed.), *From polysemy to semantic change* (pp. 163–215). Amsterdam: Benjamins.

Fraser, A. & Marcu, D. (2005). ISI's participation in the Romanian-English alignment task. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ParaText '05 (pp. 91–94). Stroudsburg, PA, USA: Association for Computational Linguistics.

Fraser, A. & Marcu, D. (2006). Semi-supervised training for statistical word alignment. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44 (pp. 769–776). Stroudsburg, PA, USA: Association for Computational Linguistics.

Fraser, A. & Marcu, D. (2007). Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3), 293–303.

Fung, P. & Church, K. W. (1994). K-vec: A new approach for aligning parallel texts. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 2*, COLING '94 (pp. 1096–1102). Stroudsburg, PA, USA: Association for Computational Linguistics.

Gal, Y. & Blunsom, P. (2013). A systematic Bayesian treatment of the IBM alignment models. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Stroudsburg, PA, USA: Association for Computational Linguistics.

Gale, W. A. & Church, K. W. (1991). Identifying word correspondence in parallel texts. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '91 (pp. 152–157). Stroudsburg, PA, USA: Association for Computational Linguistics.

Gao, J. & Johnson, M. (2008). A comparison of Bayesian estimators for unsupervised Hidden Markov Model POS taggers. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 344–352).: Association for Computational Linguistics.

Gelfand, A. E. & Smith, A. F. M. (1991). *Gibbs Sampling for Marginal Posterior Expectations*. Technical report, Department of Statistics, Stanford University.

Goldwater, S. (2007). *Nonparametric Bayesian Models of Lexical Acquisition*. PhD thesis, Brown University.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2006). Interpolating between types and tokens by estimating power-law generators. In *Advances in Neural Information Processing Systems 18* Cambridge, MA, USA: MIT Press.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2011). Producing power-law distributions and damping word frequencies with two-stage language models. *Journal of Machine Learning Research*, 12, 2335–2382.

Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg (Ed.), *Universals of Human Language* (pp. 73–113). Cambridge, Massachusetts: MIT Press.

Hammarström, H. & Borin, L. (2011). Unsupervised learning of morphology. *Computational Linguistics*, 37(2), 309–350.

Hammarström, H., Forkel, R., Haspelmath, M., & Nordhoff, S. (2014). *Glottolog 2.3*. Leipzig: Max Planck Institute for Evolutionary Anthropology. `http://glottolog.org`.

Harris, B. (1988). Bi-texts: A new concept in translation theory. *Language Monthly*, 54, 8–10.

Haspelmath, M. (2007). Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology*, 11(1), 119–132.

Hendery, R., San Roque, L., & Schapper, A. (forthcoming). Tree, firewood and fire in the languages of Sahul. An introduction. In P. Juvonen & M. Koptjevskaja-Tamm (Eds.), *Lexico-Typological Approaches to Semantic Shifts and Motivation Patterns in the Lexicon*. Berlin: De Gruyter Mouton.

Holmqvist, M. & Ahrenberg, L. (2011). A gold standard for English-Swedish word alignment. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, number 11 in NEALT Proceedings Series (pp. 106–113).

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., & Kolak, O. (2005). Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3), 311–325.

Hwa, R., Resnik, P., Weinberg, A., & Kolak, O. (2002). Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02 (pp. 392–399). Stroudsburg, PA, USA: Association for Computational Linguistics.

Johnson, M. & Goldwater, S. (2009). Improving nonparameteric bayesian inference: Experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09 (pp. 317–325). Stroudsburg, PA, USA: Association for Computational Linguistics.

Kalchbrenner, N. & Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (pp. 1700–1709).: Association for Computational Linguistics.

Kay, P. & Maffi, L. (2013a). Green and blue. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Kay, P. & Maffi, L. (2013b). Red and yellow. In M. S. Dryer & M. Haspelmath (Eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598), 671–680.

Knight, K. (2009). Bayesian inference with tears. `http://www.isi.edu/natural-language/people/bayes-with-tears.pdf`.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *The Tenth Machine Translation Summit*. Phuket, Thailand.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07 (pp. 177–180). Stroudsburg, PA, USA: Association for Computational Linguistics.

Kumar, S., Och, F. J., & Macherey, W. (2007). Improving word alignment with bridge languages. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 42–50). Prague, Czech Republic: Association for Computational Linguistics.

Källgren, G. (2006). *Manual of the Stockholm Umeå Corpus version 2.0*. Department of Linguistics, Stockholm University. Sofia Gustafson-Capková and Britt Hartmann (eds.).

Lardilleux, A. & Lepage, Y. (2009). Sampling-based multilingual alignment. In *Proceedings of the International Conference RANLP-2009* (pp. 214–218). Borovets, Bulgaria: Association for Computational Linguistics.

Lardilleux, A., Lepage, Y., & Yvon, F. (2011). The contribution of low frequencies to multilingual sub-sentential alignment: A differential associative approach. *International Journal of Advanced Intelligence*, 3(2), 189–217.

Lardilleux, A., Yvon, F., & Lepage, Y. (2012). Hierarchical sub-sentential alignment with Anymalign. In *Proceedings of the 16th EAMT Conference* (pp. 279–286). Trento, Italy.

Lewis, M. P., Simons, G. F., & Fennig, C. D. (2014). Ethnologue: Languages of the World, 17th edition. `http://www.ethnologue.com`.

Liang, P., Jordan, M. I., & Klein, D. (2010). Type-based MCMC. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10 (pp. 573–581). Stroudsburg, PA, USA: Association for Computational Linguistics.

Liang, P., Taskar, B., & Klein, D. (2006). Alignment by agreement. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, HLT-NAACL '06 (pp. 104–111). Stroudsburg, PA, USA: Association for Computational Linguistics.

Lin, D. & Cherry, C. (2003a). ProAlign: Shared task system description. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond - Volume 3*, HLT-NAACL-PARALLEL '03 (pp. 11–14). Stroudsburg, PA, USA: Association for Computational Linguistics.

Lin, D. & Cherry, C. (2003b). Word alignment with cohesion constraint. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003–short papers-Volume 2* (pp. 49–51).: Association for Computational Linguistics.

List, J.-M., Mayer, T., Terhalle, A., & Urban, M. (2014). CLICS: Database of Cross-Linguistic Colexifications. `http://clics.lingpy.org`.

Liu, Y., Liu, Q., & Lin, S. (2010). Discriminative word alignment by linear modeling. *Computational Linguistics*, 36(3), 303–339.

Loftsson, H. & Östling, R. (2013). Tagging a morphologically complex language using an averaged perceptron tagger: The case of Icelandic. In *Proceedings of the 19th Nordic Conference on Computational Linguistics (NODALIDA 2013)*, NEALT Proceedings Series (pp. 105–119). Oslo, Norway.

Lopez, A. & Resnik, P. (2005). Improved HMM alignment models for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ParaText '05 (pp. 83–86). Stroudsburg, PA, USA: Association for Computational Linguistics.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (pp. 281–297). Berkeley: University of California Press.

Martin, J., Johnson, H., Farley, B., & Maclachlan, A. (2003). Aligning and using an English-Inuktitut parallel corpus. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond - Volume 3*, HLT-NAACL-PARALLEL '03 (pp. 115–118). Stroudsburg, PA, USA: Association for Computational Linguistics.

Martin, J., Mihalcea, R., & Pedersen, T. (2005). Word alignment for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ParaText '05 (pp. 65–74). Stroudsburg, PA, USA: Association for Computational Linguistics.

Matusov, E., Zens, R., & Ney, H. (2004). Symmetric word alignments for statistical machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04 Stroudsburg, PA, USA: Association for Computational Linguistics.

Mayer, T. & Cysouw, M. (2012). Language comparison through sparse multilingual word alignment. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, EACL 2012 (pp. 54–62). Stroudsburg, PA, USA: Association for Computational Linguistics.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., & Lee, J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 92–97). Sofia, Bulgaria: Association for Computational Linguistics.

Melamed, I. D. (2000). Models of translational equivalence among words. *Computational Linguistics*, 26(2), 221–249.

Mermer, C. & Saraçlar, M. (2011). Bayesian word alignment for statistical machine translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11 (pp. 182–187). Stroudsburg, PA, USA: Association for Computational Linguistics.

Mermer, C., Saraclar, M., & Sarikaya, R. (2013). Improving statistical machine translation using Bayesian word alignment and Gibbs sampling. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5), 1090–1101.

Mesch, J., Rohdell, M., & Wallin, L. (2014). Annoterade filer för svensk teckenspråkskorpus. Version 2. `http://www.ling.su.se`.

Mesch, J. & Wallin, L. (2015). Gloss annotations in the Swedish Sign Language corpus. *International Journal of Corpus Linguistics*, 20(1), 102–120.

Mihalcea, R. & Pedersen, T. (2003). An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond - Volume 3*, HLT-NAACL-PARALLEL '03 (pp. 1–10). Stroudsburg, PA, USA: Association for Computational Linguistics.

Moore, R. C., Yih, W.-t., & Bode, A. (2006). Improved discriminative bilingual word alignment. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44 (pp. 513–520). Stroudsburg, PA, USA: Association for Computational Linguistics.

Naradowsky, J. & Toutanova, K. (2011). Unsupervised bilingual morpheme segmentation and alignment with context-rich hidden semi-markov models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 895–904). Portland, OR, USA: Association for Computational Linguistics.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics*, 31(3), 705–767.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., & Marsi, E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13, 95–135.

Och, F. J. & Ney, H. (2000). A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 2*, COLING '00 (pp. 1086–1090). Stroudsburg, PA, USA: Association for Computational Linguistics.

Och, F. J. & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 19–51.

Östling, R. (2013). Stagger: An open-source part of speech tagger for Swedish. *North European Journal of Language Technology*, 3, 1–18.

Östling, R. (2014). Bayesian word alignment for massively parallel texts. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers* (pp. 123–127). Gothenburg, Sweden: Association for Computational Linguistics.

Östling, R. (forthcoming). Studying colexification through parallel corpora. In P. Juvonen & M. Koptjevskaja-Tamm (Eds.), *Lexico-Typological Approaches to Semantic Shifts and Motivation Patterns in the Lexicon*. Berlin: De Gruyter Mouton.

Östling, R. (submitted a). A Bayesian model for joint word alignment and part-of-speech transfer.

Östling, R. (submitted b). Word order typology through multilingual word alignment.

Östling, R., Börstell, C., & Wallin, L. (2015). Enriching the Swedish Sign Language Corpus with part of speech tags using joint Bayesian word alignment and annotation transfer. In *Proceedings of the 20th Nordic Conference on Computational Linguistics (NODALIDA 2015)*. In press.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Petrov, S., Das, D., & McDonald, R. (2012). A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* Istanbul, Turkey: European Language Resources Association (ELRA).

Pitman, J. & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2), 855–900.

Resnik, P. & Hardisty, E. (2010). *Gibbs Sampling for the Uninitiated*. Technical Report LAMP-TR-153, University of Maryland, College Park.

Richey, M. (2010). The evolution of Markov Chain Monte Carlo methods. *The American Mathematical Monthly*, 117(5), 383–413.

Riley, D. & Gildea, D. (2012). Improving the IBM alignment models using variational Bayes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12 (pp. 306–310). Stroudsburg, PA, USA: Association for Computational Linguistics.

Rosenberg, A. & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 410–420). Prague, Czech Republic: Association for Computational Linguistics.

Schafer, C. & Drábek, E. F. (2005). Models for Inuktitut-English word alignment. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, ParaText '05 (pp. 79–82). Stroudsburg, PA, USA: Association for Computational Linguistics.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing* Manchester, UK.

Sjons, J. (2013). Automatic induction of word classes in Swedish Sign Language. Master's thesis, Stockholm University.

Snyder, B. & Barzilay, R. (2008). Unsupervised multilingual learning for morphological segmentation. In *Proceedings of ACL-08: HLT* (pp. 737–745). Columbus, Ohio: Association for Computational Linguistics.

Snyder, B., Naseem, T., Eisenstein, J., & Barzilay, R. (2009). Adding more languages improves unsupervised multilingual part-of-speech tagging: A Bayesian non-parametric approach. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 83–91).: Association for Computational Linguistics.

Spreyer, K. (2011). *Does It Have to Be Trees? Data-Driven Dependency Parsing with Incomplete and Noisy Training Data.* PhD thesis, University of Potsdam.

Strehl, A. & Ghosh, J. (2003). Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3, 583–617.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *arXiv preprint*, abs/1409.3215.

Täckström, O. (2013). *Predicting Linguistic Structure with Incomplete and Cross-Lingual Supervision.* PhD thesis, Uppsala University, Department of Linguistics and Philology.

Täckström, O., Das, D., Petrov, S., McDonald, R., & Nivre, J. (2013). Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1, 1–12.

Taskar, B., Lacoste-Julien, S., & Klein, D. (2005). A discriminative matching approach to word alignment. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05 (pp. 73–80). Stroudsburg, PA, USA: Association for Computational Linguistics.

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (pp. 985–992). Sydney, Australia: Association for Computational Linguistics.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 1566–1581.

Tiedemann, J. (2011). *Bitext Alignment.* Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In N. C. C. Chair), K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* Istanbul, Turkey: European Language Resources Association (ELRA).

Toutanova, K. & Galley, M. (2011). Why initialization matters for IBM model 1: Multiple optima and non-strict convexity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11 (pp. 461–466). Stroudsburg, PA, USA: Association for Computational Linguistics.

Toutanova, K., Ilhan, H. T., & Manning, C. (2002). Extensions to HMM-based statistical word alignment models. In *2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)* (pp. 87–94).

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03 (pp. 173–180). Stroudsburg, PA, USA: Association for Computational Linguistics.

Tufis, D., Ion, R., Ceausu, A., & Stefanescu, D. (2005). Combined word alignments. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts* (pp. 107–110). Ann Arbor, MI: Association for Computational Linguistics.

Urban, M. (2012). *Analyzability and Semantic Associations in Referring Expressions: A Study in Comparative Lexicology.* PhD thesis, Leiden University.

Vilar, D., Popović, M., & Ney, H. (2006). AER: Do we need to "improve" our alignments. In *In Proceedings of the International Workshop on Spoken Language Translation* (pp. 205–212).

Vogel, S., Ney, H., & Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96 (pp. 836–841). Stroudsburg, PA, USA: Association for Computational Linguistics.

Wälchli, B. & Cysouw, M. (2012). Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics*, 50(3), 671–710.

Wallin, L. (1994). *Polysyntetiska tecken i svenska teckenspråket.* PhD thesis, Stockholm University, Department of Linguistics.

Wallin, L., Mesch, J., & Nilsson, A.-L. (2014). *Transkriptionskonventioner för teckenspråkstexter (Version 5).* Technical report, Sign Language, Department of Linguistics, Stockholm University.

Wang, Z. & Zong, C. (2013). Large-scale word alignment using soft dependency cohesion constraints. *Transactions of the Association for Computational Linguistics*, 1, 291–300.

Wichmann, S., Müller, A., Wett, A., Velupillai, V., Bischoffberger, J., Brown, C. H., Holman, E. W., Sauppe, S., Molochieva, Z., Brown, P., Hammarström, H., Belyaev, O., List, J.-M., Bakker, D., Egorov, D., Urban, M., Mailhammer, R., Carrizo, A., Dryer, M. S., Korovina, E., Beck, D., Geyer, H., Epps, P., Grant, A., & Valenzuela, P. (2013). *The ASJP Database (version 16)*. Leipzig. `http://email.eva.mpg.de/~wichmann/languages.htm`.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. `http://tla.mpi.nl/tools/tla-tools/elan/`.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3), 377–403.

Wu, D. & Xia, X. (1994). Learning an English-Chinese lexicon from a parallel corpus. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas* (pp. 206–213).

Wu, H. & Wang, H. (2007). Comparative study of word alignment heuristics and phrase-based SMT. In *Proceedings of the MT Summit XI*.

Yamada, K. & Knight, K. (2001). A syntax-based statistical translation model. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics* (pp. 523–530). Toulouse, France: Association for Computational Linguistics.

Yarowsky, D. & Ngai, G. (2001). Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01 (pp. 1–8). Stroudsburg, PA, USA: Association for Computational Linguistics.

Yarowsky, D., Ngai, G., & Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01 (pp. 1–8). Stroudsburg, PA, USA: Association for Computational Linguistics.

Zhang, H. & Gildea, D. (2004). Syntax-based alignment: Supervised or unsupervised? In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04 (pp. 418–424). Geneva, Switzerland: COLING.

Zhao, S. & Gildea, D. (2010). A fast fertility Hidden Markov Model for word alignment using MCMC. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 596–605). Cambridge, MA, USA: Association for Computational Linguistics.